S1. Further details of the biomass and abundance models

a. Restrictions on the checklists

Besides those mentioned in the main text, we also imposed the following additional restrictions:

- Checklists had to lie inland. This was determined using the world map from the Natural Earth project, as implemented by the maps package in R. To allow for inaccuracy in the coastline polygons, an additional 50 km buffer was added to the polygons.
- For observation protocol, only standard travelling or stationary counts.
- Number of observers ≤ 10.
- Observation duration ≤ 5 hrs.
- Distance travelled during observation ≤ 5 km.
- $-4 \leq$ checklist calibration index ≤ 4.

The latter four restrictions were meant to avoid extreme values of the sampling effort predictors, for various reasons. For example, topographical and land cover predictors were derived at a certain spatial resolution or neighbourhood size, and hence may not adequately represent a checklist where the distance travelled greatly exceeded these scales. Also, we noticed that as we increased observation duration beyond five hours, instead of shifting towards higher values, the distribution of detected biomass started shifting towards lower values. This was true even after we had controlled for seasonality, hence ruling out a possible explanation that the long-duration checklists were dominated by submissions from the Christmas Bird Count during winter, a period of low biomass. One possible reason could be that observers reporting long observation durations were more likely to have included long breaks between sampling bouts as part of the duration. Therefore, excluding the long-duration checklists should improve data quality.

We also note that by restricting to "complete checklists", most checklists based on the nocturnal flight count (NFC) protocol would have been filtered out, since the protocol required that these checklists be marked as "incomplete". Therefore, most checklists (98%) were based on diurnal observations.

b. Stixel generation

Stixels were generated using the following procedure. We created 21-day windows moving in steps of 1 day, with the first useful window spanning Days -19 – 1 (supporting Day 1), and the last useful window spanning Days 366 – 386 (supporting Day 366). Here, Days that were below 1 or above 365 (or 366 in leap years) wrapped around to the previous or next year; e.g. Day 0 of Year 2016 corresponded to Day 365 of Year 2015. For each moving window, we partitioned the study area using a 750 km × 750 km randomly oriented spatial grid. Hence each grid cell together with the 21-day window defined a stixel. The partitioning was done 5 times for each 21-day window. Therefore, each spatio-temporal point fell within exactly $21 \times 5 = 105$ overlapping stixels: 21 windows with 5 spatial partitions per window. We used Albers equal-area projection with standard parallels at latitudes 28⅔° N and 63⅓° N, and central longitude 111° W, based on the WGS 84 ellipsoid.

Since data from Year 2017 were not readily available during our analysis, this meant that windows starting on Day 347 or later would have less data than expected, due to these windows requiring data from Day 367 of Year 2016 or later. To avoid this issue, we redefined a calendar year to begin 20 days earlier. For example, "Year 2016" was defined to begin on 2015/12/12, and ended on 2016/12/12, so that even the last useful window starting on Day 366 would not require data from Year 2017 since Day 386 of "Year 2016" now corresponded to 2016/12/31. On a related note, even though we intended our data to begin on "Year 2007", the need to include data from Day -19 of "Year 2007" was why we included checklists starting from 2006/11/22, rather than 2007/01/01.

c. Ensemble support

Recall that in the STEM framework, the estimate at any spatio-temporal point of interest is obtained by averaging across the predictions of all base models whose stixels the point had fallen within. We define ensemble support as the number of base models predictions used in the average. As explained in Supp. Section 1b, each spatio-temporal point fell within 105 stixels, so the maximum ensemble support is 105. However, the actual ensemble support could be less because we fitted a base model to a stixel only when the number of checklists within the stixel was at least 1000 (after spatio-temporal subsampling; see Supp. Section 1e), so not every stixel had a base model that could contribute to the ensemble. We required an ensemble of at least 20 base models before the model estimate at a point was considered acceptable.

The above requirement meant that the study area had to be further restricted to regions with sufficient data. While in principle we could still allow the spatial extent of the study area to vary across the annual cycle, the temporal correlation analyses to be conducted later required year-round predictions of biomass and abundance. Hence, the study area only included regions with sufficient data *year round*. This led to conspicuous gaps in data-sparse regions such as the Northern Great Plains. We addressed this issue by using adaptive spatial sizing of stixels (Fink *et al.*, 2013), which we explain in Supp. Section 1d.

d. Adaptive design

The uneven spatial density of eBird checklists means that under a fixed stixel design, data-sparse regions may not have sufficient ensemble support since many stixels do not meet the minimum number of checklists required to fit a base model. Adaptive stixel design (Fink *et al.*, 2013) based on QuadTrees partitioning (Samet, 1984) were designed to address this issue by allowing larger stixels in data-sparse regions.

For our STEM model, we wanted a simpler adaptive design that involved just a slight modification of the uniform grid. One way to do so was to double the spatial dimensions of any grid cell that did not meet the minimum requirement under the original dimensions, so that more checklists could be included. However, this enlarged grid cell would then overlap with its neighbours, and hence increase the maximum possible ensemble support for points within the overlapping areas, which we did not want. Hence, instead of a truly adaptive stixel design, we only implemented an adaptive "checklist inclusion" design, in that stixels could be enlarged for the purpose of including more checklists, but the base model of an enlarged stixel was still only used to generate predictions over the original stixel before enlargement.

In the context of bias-variance tradeoff, enlarging a stixel meant lower variance at a cost of higher bias, especially if the number of external checklists (checklists included only after enlargement) greatly exceeded the internal ones (checklists included in the original stixel). Therefore, we implemented the following ad-hoc rules:

- If number of internal checklists $\geq$ 1000, adaptive inclusion was not required.

- If 500 ≤ number of internal checklists < 1000, maximum number of external checklists included = number of internal checklists. (If there were more external checklists than the maximum, they were subsampled down to the maximum.)
- If 334 ≤ number of internal checklists < 500, maximum number of external checklists included = number of internal checklists × 2.
- If 250 ≤ number of internal checklists < 334, maximum number of external checklists included = number of internal checklists × 3.
- If number of internal checklists < 250, adaptive inclusion was not implemented.

The purpose of these rules was to allow the total number of checklists to reach at least 1000, but yet limit the external : internal ratio as much as possible, with the worst case scenario being arbitrarily set at 3 : 1.

e. Spatio-temporal subsampling

Within stixels, uneven sampling of eBird data across space and time can affect the quality of predictions. To address this, we performed spatio-temporal subsampling (Robinson et al., 2018; Fink et al., 2020a; Johnston et al., 2021). Before fitting each base model, we partitioned each stixel into a 5 km × 5 km × 7 day spatio-temporal grid, and one checklist was randomly sampled per grid cell.

f. Interannual data balancing

The increase in the rate of checklist submissions over the 10-year study period meant that there were many more checklists from later than earlier years. This interannual imbalance could affect model performance in the earlier years. To address this imbalance, we oversampled checklists from earlier years and subsampled those from later years until the number of checklists each year was at the pre-sampling rounded average (Fink *et al.*, 2020). A small jitter was added to the predictor values of each oversample; the jitter was randomly sampled from a uniform distribution with bounds [-0.05, 0.05] × standard deviation of the predictor values available in the stixel training set.

However, we also wanted to avoid a scenario where, say, there was only one checklist from 2007 which we then had to resample 100 times just to reach the average. Therefore, we modified the notion of a stixel to also include year and not just location and day: a year was considered to have fallen within a stixel only if the stixel contained at least 20 checklists from that year (after spatio-temporal subsampling). The number 20 was arbitrarily chosen to limit the amount of oversampling required. Base models were fitted using balanced checklists only from years that fell within the stixel, and only provided ensemble support for the same years. Hence, ensemble support at the same location and day could still vary by year in data-sparse regions, since base models in those regions needed not support every year in the study period.

g. Enhancing interannual contrasts in model predictions

We found that even after balancing the data across years, the model tended to predict similar-looking biomass distributions the same day across different years. To enhance the contrasts, first we set year as a categorical rather than a continuous predictor, since there was no reason to require that the split at any node of a regression tree keep consecutive years together. For example, spring migration could be delayed by weather conditions in one year but not in the years immediately before or after. Second, we forced the regression trees in the random forest to always consider year as a possible predictor at each node.

To verify that the enhanced interannual contrasts resulting from the above were realistic, we compared the distribution maps to those obtained from "one-year" biomass models, during peak spring migration in the northeast. Each "one-year'' model was fitted using only checklists from one year, and used to generate predictions only for that year. Hence, in the context of bias-variance tradeoff, predictions from the "one-year" models were expected to have more variance (since each model was fitted with less data), but less bias (since checklists from other years were not included). We found that differences in migration phenology suggested by the enhanced contrasts were indeed present in the "one-year" models and hence meaningful.

h. Train-test split

We split the data into a training set and a test set in the ratio 80:20. For model evaluation to be accurate, the test set should ideally be independent of the training set. However, spatio-temporal autocorrelations are likely to be present in the data. We illustrate this issue using winter finches, even though they are not included among the nocturnal migrant species. If a flock of winter finches moves to a new foraging area, the same flock might remain there for a few weeks, so checklists from that area and period are likely to report comparable counts, especially if there is an influx of observers specifically looking for rare species within the flock. Therefore, the counts are autocorrelated and do not reflect the true underlying variation in winter finch abundance.

To address this, we partitioned the entire region of study using a 5 km × 5 km spatial grid. Each year, 80% of the grid cells were designated as training cells, and 20% as test cells. All checklists in the training cells were assigned to the training set, and those in the test cells assigned to the test set. Therefore, test data could share the same grid cells with training data from different years, but not the same year. This procedure ensured that the test set was independent of the training set, provided that spatio-temporal autocorrelations decayed within a lengthscale of a few km and a timescale of a year. The decision to allow for such a long timescale was made with the breeding season in mind, where individuals from large-bodied species might hold territories for multiple months.

i. Spatial subsampling in the test set

To avoid having checklists from highly-sampled locations dominate the calculations of model performance, within each 7-day window we first partitioned the study area into a 5 × 5 km grid. We then subsampled one checklist per grid cell and calculated the percentage variance in the subsample explained by the model. So as not to waste the rest of the test set, we repeated the subsampling and calculation a total of 30 times, and then averaged the percentage variance explained across subsamples.

S2. Spatial correlation analysis based on the seasonal EVI difference

We calculated the seasonal EVI difference using the following procedure. At each site (grid location) and week, we averaged the EVI across the years of study. We then constructed the weekly EVI trajectory at the site, averaged across years. The minimum of this averaged trajectory was then taken to be the site minimum. We then subtracted the (non-averaged) weekly EVI trajectory each year by this site minimum to obtain the EVI difference used in the spatial correlation analyses. The reason for averaging across years when determining the site minimum is because a small but nonetheless substantial number of sites had weekly trajectories that were highly variable between years. Note that since the minimum along a non-averaged weekly trajectory may be lower than the minimum calculated this way, it is possible to obtain a negative EVI difference.

Supp. Figure S2 shows the averaged EVI trajectories at 10,000 randomly-selected sites. Each trajectory has been scaled so that the EVI values span 0 to 1. We see that between December and March, most sites had values that were close to the site minimum 0, so the EVI difference would be low throughout most of the study area. As a result, spatial analyses based on EVI difference could be misleading for two reasons. First, the EVI difference may no longer have the intended ecological meaning. Based on the conceptual model in Hurlbert & Haskell (2003), the site minimum roughly represents resource usage by the year-round residents, so subtracting by the site minimum gives the surplus available to the migrants. However, this is clearly a crude approximation; only when the EVI difference is large enough can we ignore the uncertainties in this approximation, which is not the case during the winter period. Second, any small but systematic errors could have an outsized effect when most values are already low. Therefore, we only calculated the spatial correlation coefficients each week in a half-year period spanning May to September, when most grid locations have EVI values substantially higher than the site minima.

S3. Additional results from modelling the regional patterns of temporal associations using bioclimatic variables

a. Variable importance based on forward stepwise selection

As mentioned in the main text, model performance was 92% when the two most important bioclimatic variables, BIO11 (mean temperature of coldest quarter and BIO18 (precipitation of warmest quarter), were included. Model performance saturated at 98% with the addition of BIO4 (temperature seasonality) and BIO19 (precipitation of coldest quarter); further variable additions only led to a cumulative improvement of less than 1%. See Supp. Figure S9.

Note that since many bioclimatic variables were highly correlated with one another, model performance could sometimes remain high when a previously selected variable was replaced by an unselected highly-correlated counterpart. For example, BIO1 (annual mean temperature) was highly correlated with BIO11 (Spearman's $\rho = 0.97$), and a model with BIO1 + BIO18 + BIO4 + BIO19 performed just as well (98%) as the one with BIO11 + BIO18 + BIO4 + BIO19.

b. Permutation variable importance

Four variables had noticeably higher permutation importance than the rest: BIO11 (mean temperature of coldest quarter), BIO1 (annual mean temperature), BIO6 (minimum temperature of coldest month) and BIO10 (mean temperature of warmest quarter), see Supp. Figure S8(b). Note, however, that all four variables were highly correlated with one another (Pearson correlation coefficient: 0.97 for BIO11-BIO1, 0.99 for BIO11-BIO6, 0.82 for BIO11-BIO10, 0.94 for BIO1-BIO6, 0.93 for BIO1-BIO10, and 0.78 for BIO6-BIO10), so they are likely to contain redundant information from the point of view of forward stepwise selection. Indeed, the partial dependence plots of BIO1, BIO6 and BIO11 had nearly identical shapes (Supp. Figure S10).

Supplementary figures

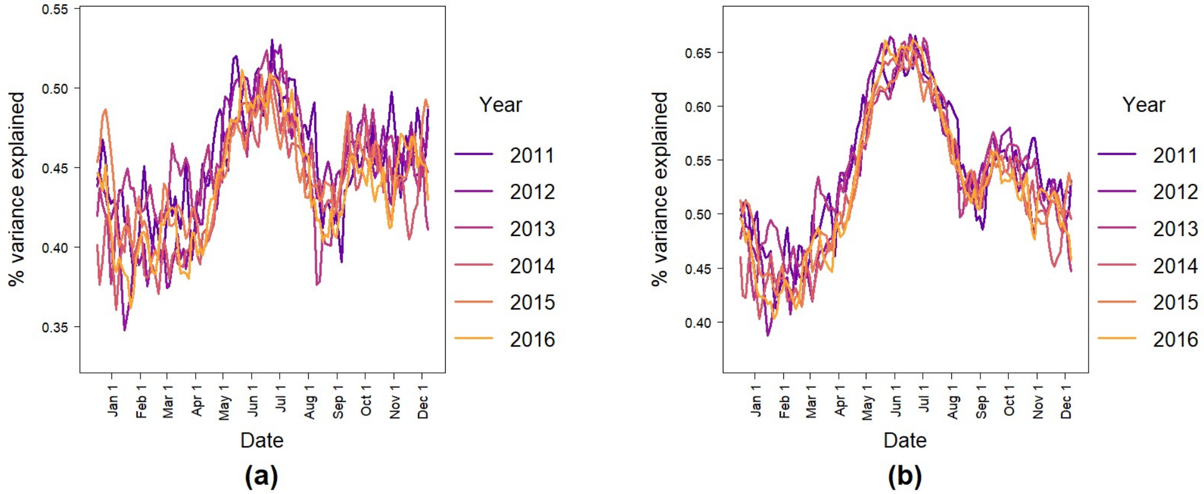


Figure S1 | Predictive performance of the STEM models. (a) Biomass model. **(b)** Abundance model. Each curve represents the model performance on holdout test data from one year, defined as the percentage variance explained by the model in a 7-day moving window.

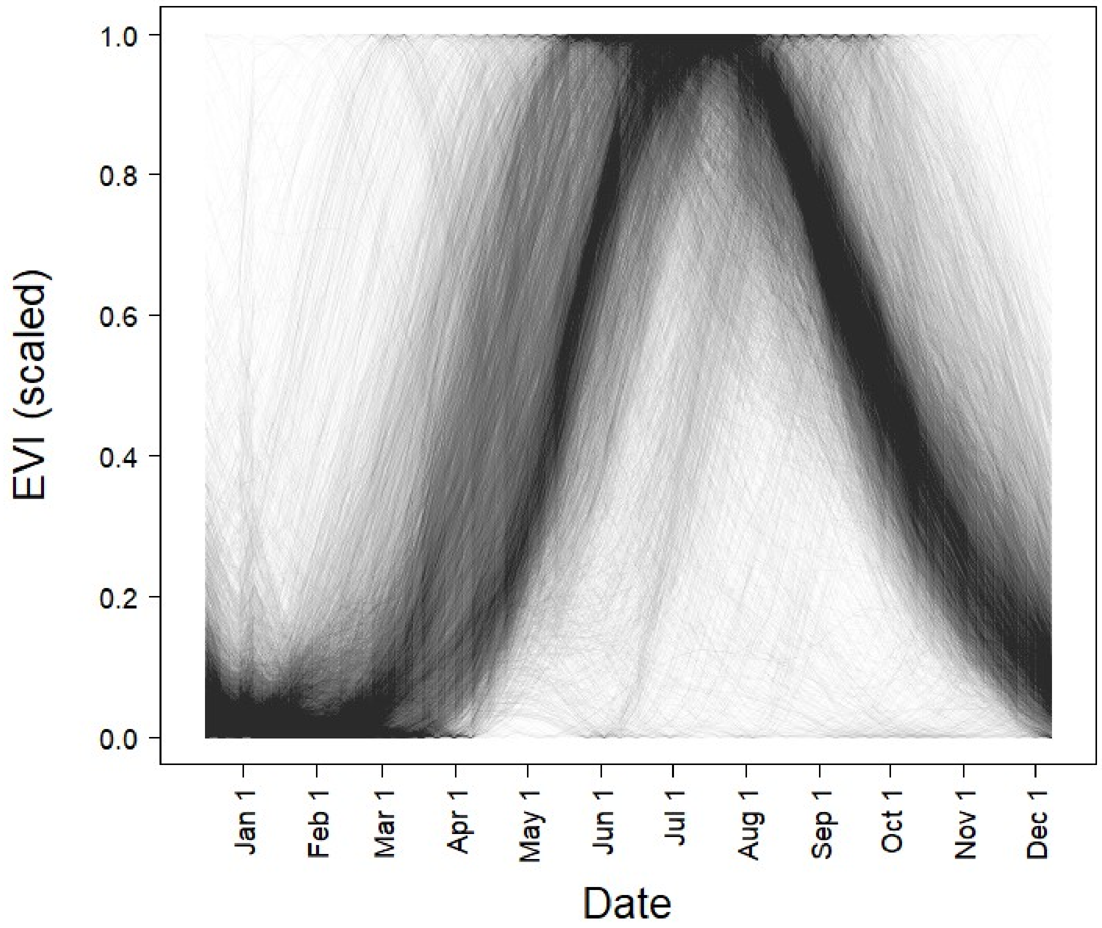
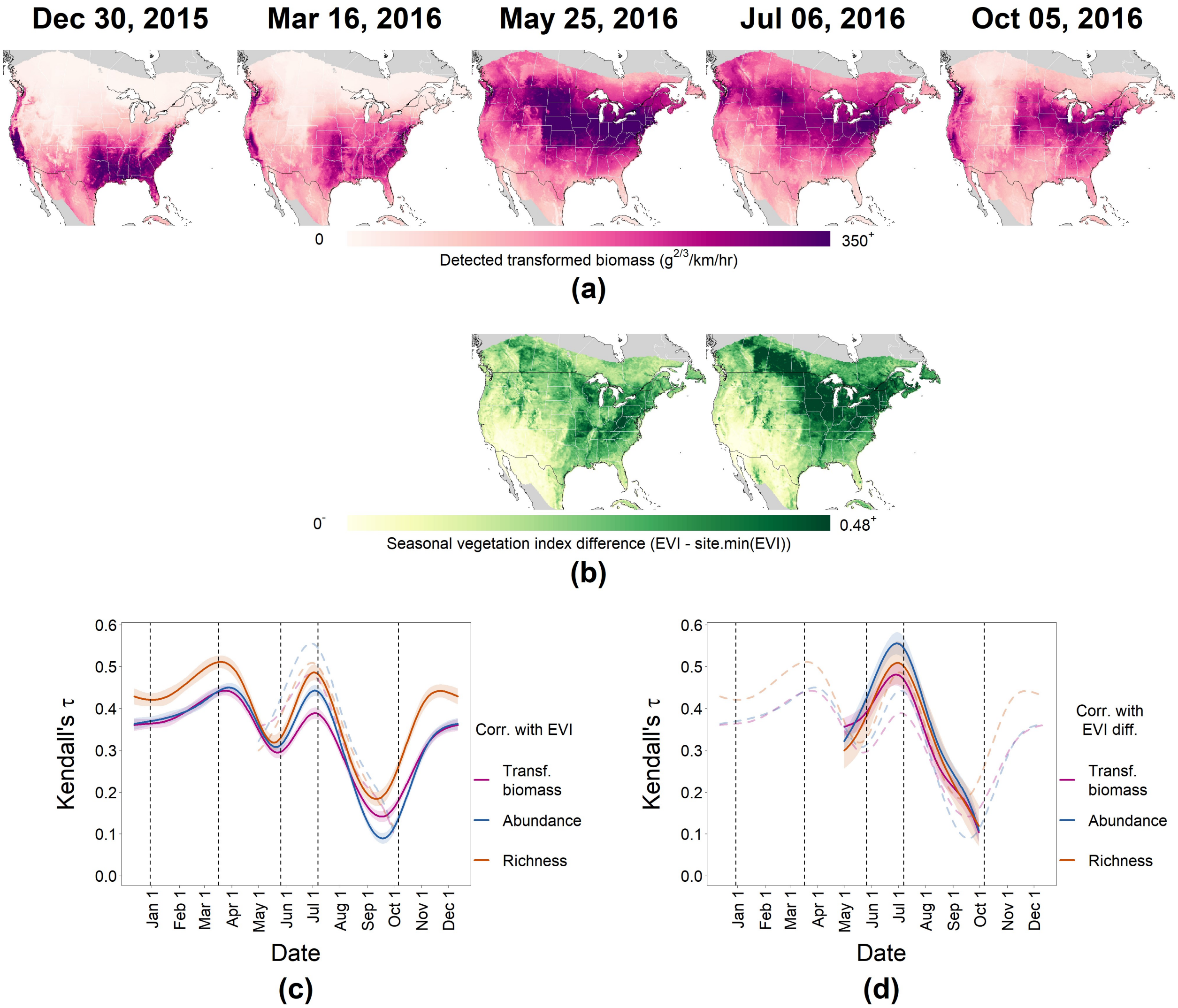


Figure S2 | Weekly EVI trajectories at 10,000 randomly chosen grid locations. Each trajectory was averaged across the years of study, and then scaled so that the values span 0 to 1.



(a)

(b)

(c)

(d)

Figure S3 | Additional maps and spatial associations. (a) Spatial distributions of total transformed biomass estimated at the same five dates in Figure 1, defined as the sum of (body mass)$^{2/3}$ across individuals. The transformed biomass is motivated by empirical allometric relationships between metabolic rate and body size for birds, which often found a scaling exponent close to 2/3. **(b)** Seasonal differences in primary productivity, defined as the difference between EVI and the site minimum. This is a crude approximation of the surplus productivity available to migratory birds. Note that we only evaluated the EVI difference between May and September, for reasons explained in Supp. Section S2. **(c)** and **(d)** are similar to Figures 1(e) and (f), except that total biomass has been replaced by the total transformed biomass, .

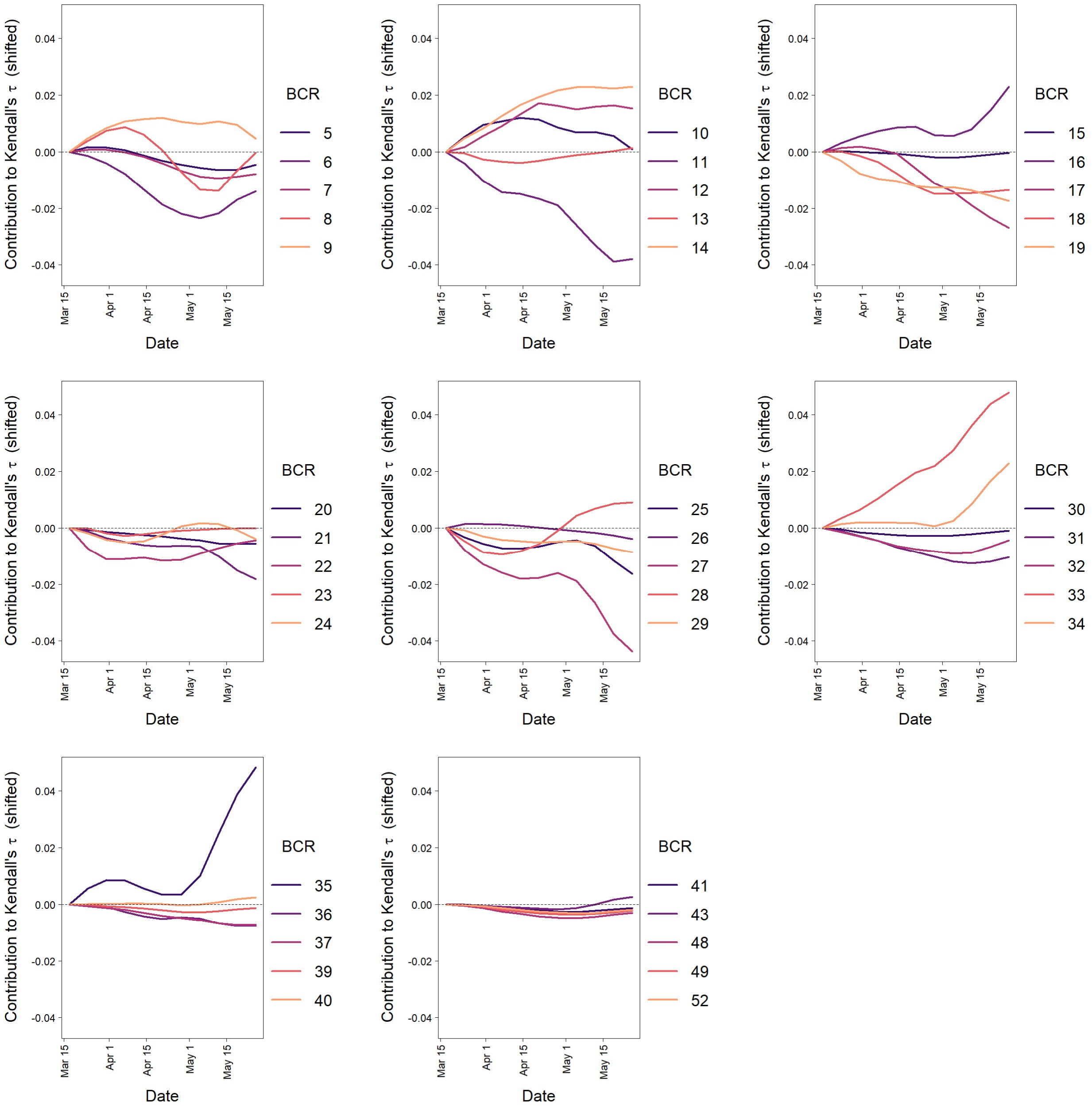


Figure S4 | Explaining the decrease in spatial association between biomass and EVI from early to late spring. We estimated the contribution of each Bird Conservation Region (BCR; Sauer *et al.*, 2003) to the spatial correlation coefficient each week, by calculating the (negative of the) change in the coefficient after removing paired data points from all grid locations within the BCR. In this figure, we shifted the contributions so that they started at zero in early spring; this makes it easier to see how the contributions changed between early and late spring. The contributions decreased by the greatest amount in BCRs 6, 11, 17, 18, 19, 21, 25 and 27. These BCRs mostly lie in the Northern Great Plains or in Southeastern USA. Hence the decrease in spatial association is primarily driven by biomass increasing in the Northern Great Plains even when the EVI remained relatively low, as well as biomass decreasing in Southeastern USA even when EVI remained relatively high.

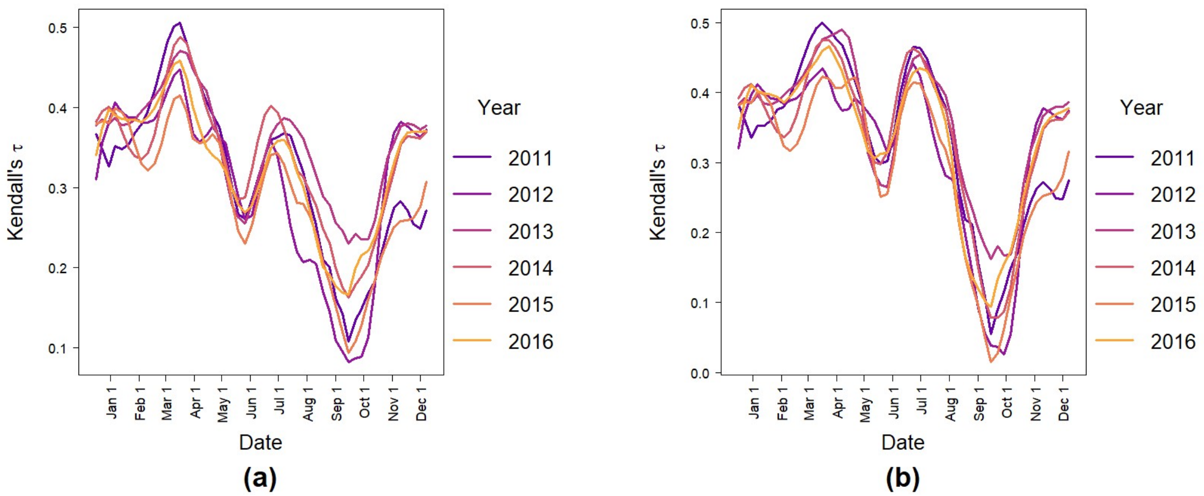


Figure S5 | Spatial associations between EVI and biomass or abundance. (a) Biomass. **(b)** Abundance. Each curve represents the spatial cross-correlation coefficients evaluated weekly across one year, without any smoothing applied. We see that each set of curves show the same seasonal variations as the corresponding GAM fit in Figure 1(e).

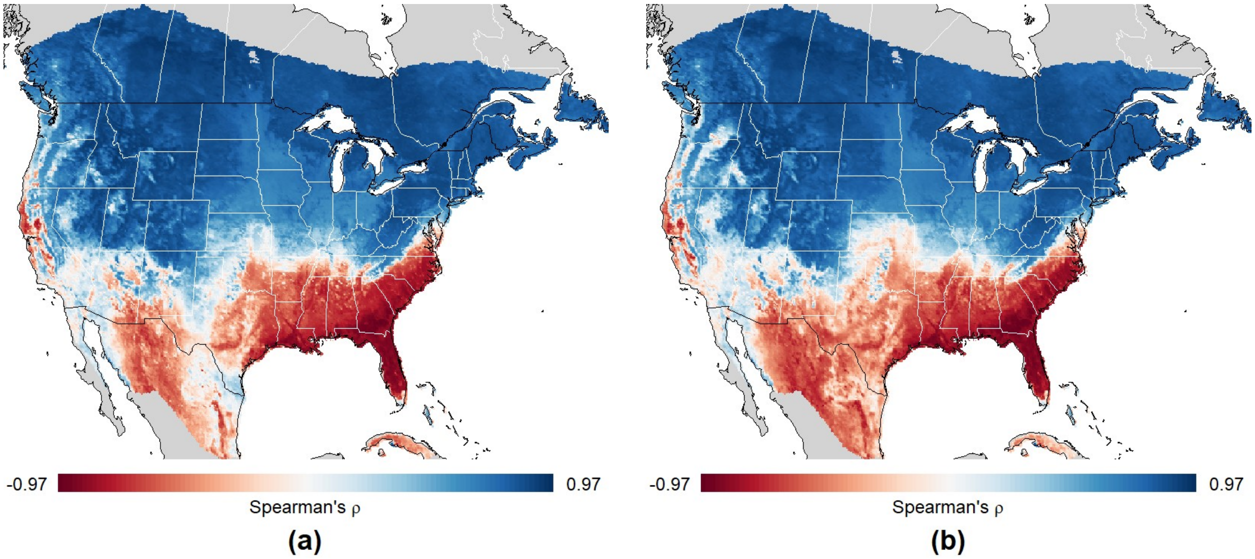


Figure S6 | Association between primary productivity and the avian biomass and abundance. Similar to Figure 2(a) and Supp. Figure S5(a), except with Spearman instead of Kendall correlation coefficients. **(a)** Biomass. **(b)** Abundance.

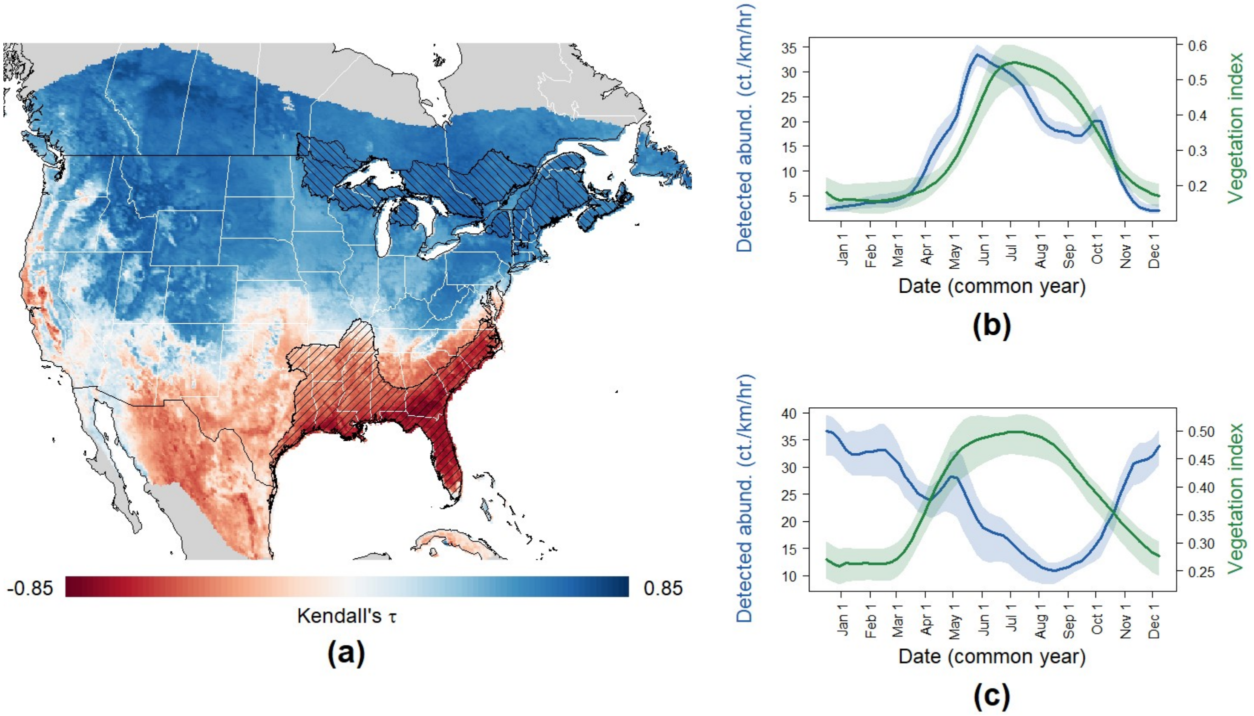


Figure S7 | Temporal association between detected abundance and primary productivity. Equivalent to Figure 2 except with abundance as the avian response instead of biomass.

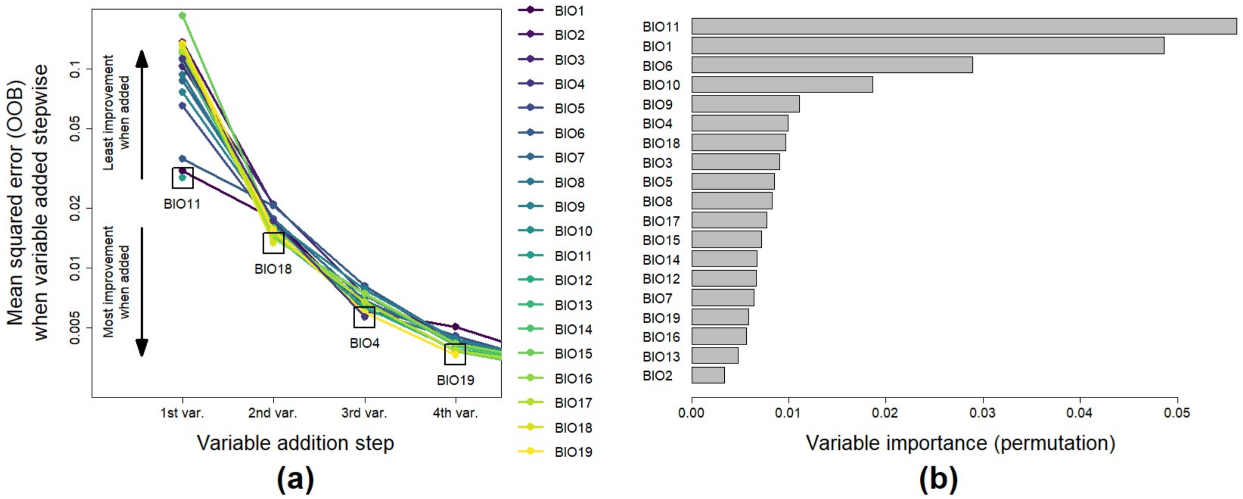


Figure S8 | Ranking importance of bioclimatic variables in predicting the temporal correlation between productivity and biomass. (a) Using forward stepwise addition. The top four variables were BIO11 (mean temperature of coldest quarter), BIO18 (precipitation of warmest quarter), BIO4 (temperature seasonality), and BIO19 (precipitation of coldest quarter). **(b)** Using permutation importance. The top four variables were BIO11 (mean temperature of coldest quarter), BIO1 (annual mean temperature), BIO6 (minimum temperature of coldest month) and BIO10 (mean temperature of warmest quarter).

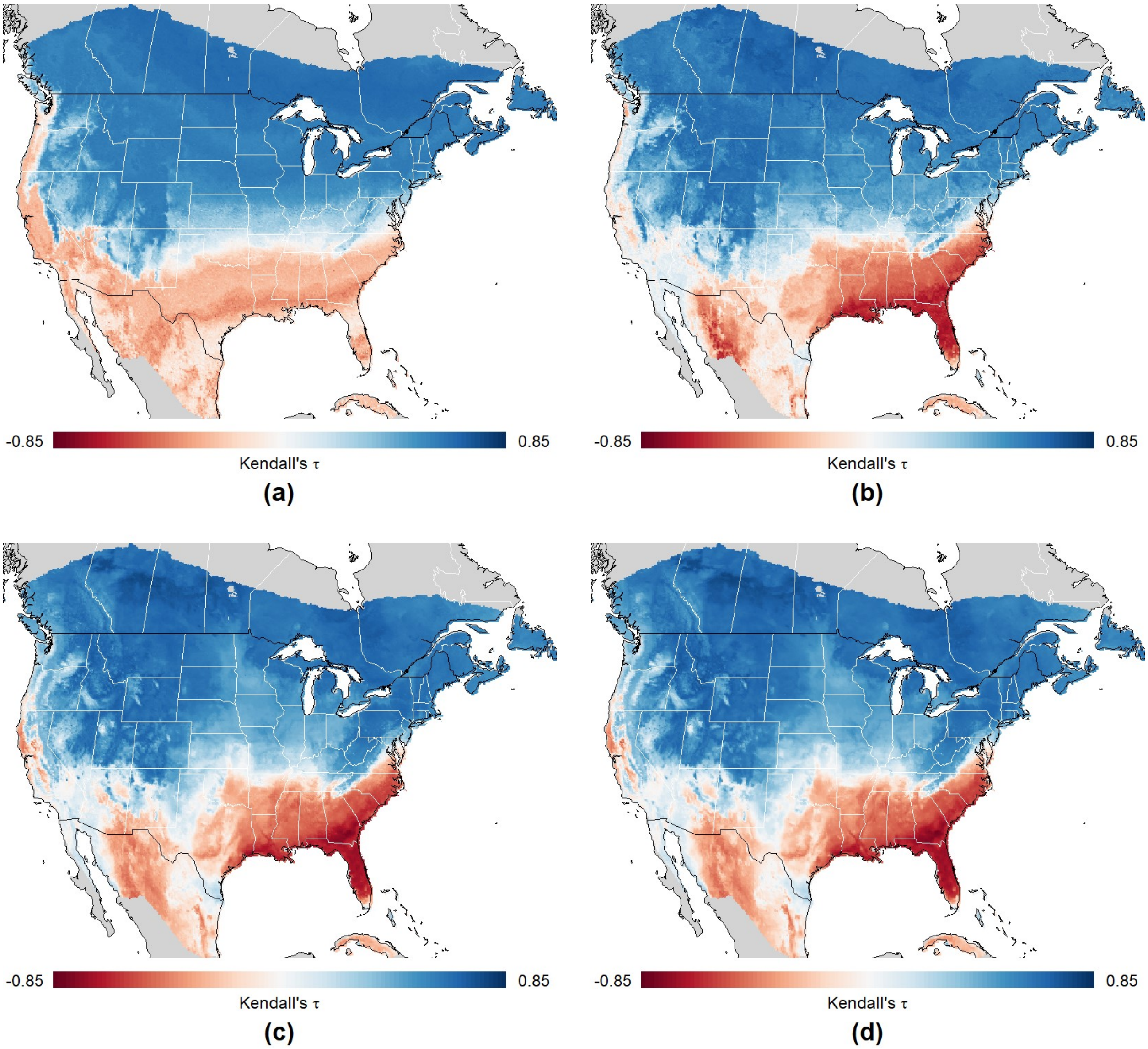


Figure S9 | Predicting the temporal correlations using bioclimatic variables, chosen by stepwise selection. (a) One-variable random forest model with BIO11. **(b)** Two-variable model with BIO11 and BIO18. **(c)** Four-variable model with BIO11, BIO18, BIO4 and BIO19. **(d)** Model using all 19 bioclimatic variables. Notice that the four-variable model is already sufficient to reproduce all the important features in Figure 2(a).

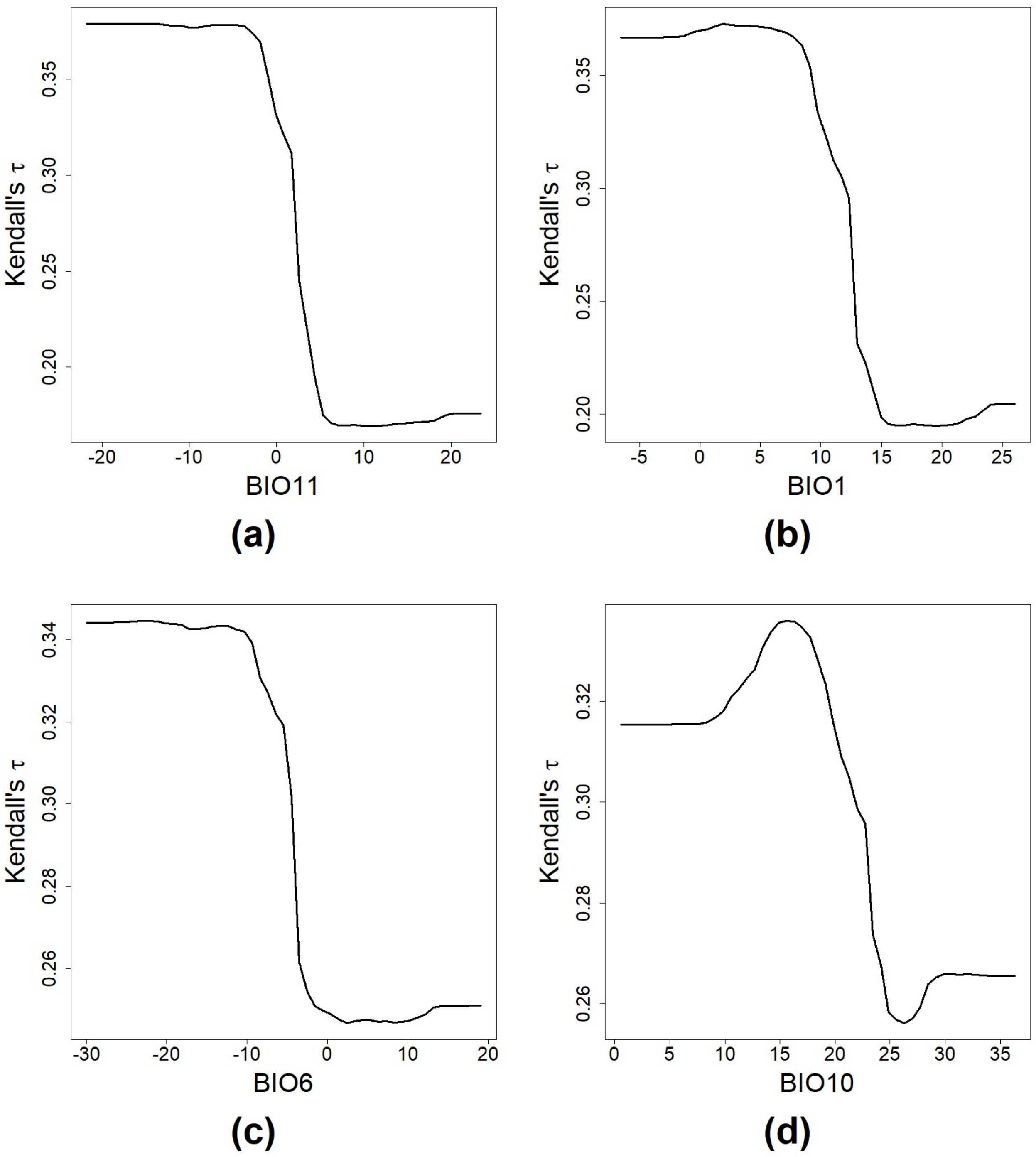


Figure S10 | Partial dependence plots showing the effects of bioclimatic variables on the temporal cross-correlation coefficients. Here, we show the plots for the four variables with the highest permutation importance. **(a)** BIO11, the mean temperature of coldest quarter, **(b)** BIO1, the annual mean temperature, **(c)** BIO6, the minimum temperature of coldest month, and **(d)** BIO10, the mean temperature of warmest quarter. Notice that they are nearly identical since the four variables are very highly correlated.

Supplementary references

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W.M. & Kelling, S. (2020) Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, **30**, e02056.

Fink, D., Damoulas, T. & Dave, J. (2013) Adaptive spatio-temporal exploratory models: Hemisphere-wide species distributions from massively crowdsourced ebird data. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* AAAI'13., pp. 1284–1290. AAAI Press, Bellevue, Washington.

Hurlbert, A.H. & Haskell, J.P. (2003) The Effect of Energy and Seasonality on Avian Species Richness and Community Composition. *The American Naturalist*, **161**, 83–97.

Samet, H. (1984) The Quadtree and Related Hierarchical Data Structures. *ACM Computing Surveys*, **16**, 187–260.

Sauer, J.R., Fallon, J.E. & Johnson, R. (2003) Use of North American Breeding Bird Survey Data to Estimate Population Change for Bird Conservation Regions. *The Journal of Wildlife Management*, **67**, 372–389.