Pennsylvania
Department of
Environmental Protection

OFFICE OF WATER PROGRAMS

BUREAU OF CLEAN WATER

STATE SCALE STATISTICAL SURVEY TECHNICAL DEVELOPMENT REPORT

2026

Prepared by:

Dustin Shull and Matthew Shank
Pennsylvania Department of Environmental Protection
Office of Water Programs
Bureau of Clean Water
11th Floor: Rachel Carson State Office Building
Harrisburg, PA 17105

2026

ACKNOWLEDGEMENTS

Many individuals contributed their expertise and efforts towards data collection, sample identification, and conceptual scoping of this project. Their involvement greatly improved this project, and the understanding of biological integrity across Pennsylvania. We also greatly appreciate the guidance and technical review by Michael Dumelle, Leah Ettema, Greg Pond, Nicole Sadecky, and Garrett Stillings at the United States Environmental Protection Agency.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The Pennsylvania Department of Environmental Protection (DEP) applied modeling techniques that used measures of biological integrity of sampled stream sites and corresponding land cover to estimate biological integrity in unsampled stream reaches. The performance of the model enabled the DEP to generate estimates of the extent of streams in good, fair, poor and very poor condition and report on water quality statewide. Results suggest that while good and very poor condition streams are staying consistent over the past 20 years, there is likely a 5% increase in fair condition and a 5% decrease in poor condition streams. These condition estimates provide valuable insights to inform prioritization of actions to protect and restore water quality.

INTRODUCTION

State agencies such as the DEP rely on the award of Federal Monitoring Initiative Funds under Clean Water Act Section 106 (106 Supplemental Grant) to fill resource gaps and enhance their water quality monitoring program. The 106 Supplemental Grant requires states, interstate agencies, and tribes to conduct specific activities to receive these funds including implementation of comprehensive monitoring strategies to include state-scale statistically valid surveys of water condition and collaboration on statistically valid surveys of the nation's waters. These state-scale statistically valid surveys must adhere to the following criteria (USEPA 2008):

1) State is implementing a statistical survey design that provides condition estimates for a population of waters (e.g., streams, rivers, lakes, coastal waters, or wetlands) of the state based on an unbiased, representative sample of a subset of those waters.
   a) The state assesses water quality conditions using core indicators for at least one designated use consistent with the Elements of a State Water Monitoring and Assessment Program guidance. Over time, state surveys incorporate a full suite of appropriate biological, chemical and physical indicators as described in the guidance. Initial statistically valid, probability surveys (through 2012), however, may be based on a subset of indicators tailored to specific water quality issues (e.g., biological integrity, recreation, fish consumption, etc.).
   b) The implementation of a state-scale statistically valid survey may span several years. A state may use a rotating basin approach and survey different watersheds over time, or spread the sites required across the state over multiple years—if these surveys can be aggregated for a state-scale survey. For example, a state may choose to sample 10 sites each year over a five-year period.
   c) States may use methods and protocols employed in the national surveys, or state methods.
   d) State surveys aim to achieve 90% confidence ±10%. This typically requires about 50 sites.
   e) Surveys assess at least one water type (streams, lakes, rivers, coastal waters, or wetlands).
   f) A state's monitoring strategy indicates a commitment to continuing statewide statistical surveys as a component of its comprehensive monitoring program.
2) State continues to participate in the national/regional scale surveys, unless the state-scale survey is fully consistent with national survey design and methods.

3) State reports the results of the state scale survey by 2012, preferably as a component of the state's Integrated Report/305b/303d (narrative form) and/or in the probability survey module of the Assessment Database. (EPA will modify this module to accommodate state assessment categories, e.g., good/ fair/poor, biocondition gradient levels, etc.).

These requirements are meant to improve state and tribal capacity to monitor and report on water quality condition and trends, and they are most often fulfilled by implementing a probabilistic survey design, identical or like the survey design used by the United States Environmental Protection Agency's (USEPA) National Aquatic Resource Surveys (spsurvey R package, Dumelle et al. 2023a, Nahlik et al. 2025). DEP, however, conducts mostly targeted survey design with minimal probabilistic survey design added for specific purposes. For example, bacteria surveys for the assessment of the Water Contact Sports use in Pennsylvania are mostly targeted, but DEP conducts stratified probabilistic sampling using a rotating basin approach to meet 106 Supplemental Grant requirements. One of the major downsides to stratified probabilistic sampling using a rotating basin approach is that it can take many years to build enough information to develop a state-scale statistically valid water quality condition for a particular protected use or stressor. As a result, it would be beneficial to DEP if there was a way to utilize its large dataset collected using a targeted sampling design while also meeting the federal requirements of state-scale statistically valid surveys.

Targeted and probabilistic survey designs are fundamentally different survey types with distinct goals and limitations. For example, a probabilistic survey is designed for statistical inference about a larger population, while a targeted design is typically representative of the local area or stream reach. However, DEP employs a judgement-based assessment strategy that uses key concepts of representativeness (such as land cover changes and major water quality influences) to delineate decisions beyond the stream segment level (Shull and Pulket 2023). As a result, DEP has amassed a sufficiently robust targeted dataset that covers entire state, includes samples throughout the hydrologic network, and represents a gradient of all land covers found in PA. This dataset has attribution to the National Hydrography Dataset (NHD) stream network, which allows relationships with land cover composition to be established and predictions to be made to unmeasured locations. These predictions can be averaged and used to make statistical inferences about a larger population. This is generally considered a model-based approach. Rather than relying on a probability sample (as in a design-based approach), model-based approaches leverage spatial autocorrelation (i.e., dependence) in the data to inform a model structure (Cressie 1993, Zimmerman and Ver Hoef 2024, Yang and Jin 2010, Yang et al. 2017, Zaresefat et al. 2024, Zhang and Du 2019). These types of spatial models are enhanced by including co-variates that are directly related to the data at hand. In particular, the freely available R package spmodel builds upon commonly used R functions like lm() to provide tools that incorporate spatial dependence into statistical models (Dumelle et al. 2023a). These models can be used to predict the average in a region – the same quantity estimated via the population-based random sampling (i.e., design-based) approach. Various studies have shown that as long as these spatial models are properly specified, they often yield estimates of averages that are more precise (i.e., lower standard errors) than similar quantities derived from design-based approaches (Ver Hoef 2008, Dumelle et al. 2022).

The intended goal of this work is to determine whether a combination of data from a targeted survey design along with a model-based approach can produce a valid biological-based water quality condition (i.e., biological condition) for Pennsylvania while also meeting all the grant requirements listed above. If successful, this work can potentially be used for other water quality condition indicators that meet representativeness requirements in Pennsylvania, and as a model for other jurisdictions that primarily implement targeted sampling to meet the requirements of 106 Supplemental Grant funding.

METHODS

The analytical approach utilized to model biological condition across Pennsylvania involved numerous steps and was highly tailored to DEP data due to programmatic data management and assessment strategies. This approach involved compiling data, spatial linear modeling, and indicator block-Kriging. The flow chart (Figure 1) represents a generalized approach to this analytical framework but does not follow the exact steps used in this study. The specific steps used are described in the Methods section herein; Figure 1 is a generalization intended to assist other entities that are interested in applying this analytical framework to their data.

Throughout this report, the terms observed and prediction datasets are used. Observed data were macroinvertebrate samples with Index of Biotic Integrity (IBI) scores. Prediction data were all National Hydrography Dataset (NHD) flowlines representing the target population where biological condition was to be predicted.

The generalized analytical approach involved three steps. See Figure 1 for a graphical representation.

- First, sample data were compiled across a geographic area. In this case, macroinvertebrate IBI scores were compiled from available samples (observed data) that were associated with NHD flowlines. Land cover covariates were accessed from StreamCat (Hill et al. 2016) for all NHD flowlines in Pennsylvania, representing the target population (prediction data). NHD flowlines were converted to points using the centroid of each reachcode.
- The second step used the splm function in the spmodel R package (Dumelle et al. 2023b) to create a model of IBI score, including land cover covariates using observed data. Kriging was then performed on the prediction dataset to predict IBI score at all reachcode centroids in the target population. The predictions were then translated back to the NHD flowlines. This was an optional step, but helpful for visualization purposes.
    - This step provided point estimates of IBI score for each centroid in the prediction dataset, but did not provide overall estimates of stream condition across Pennsylvania or error estimation.
- The third step, termed indicator block-Kriging produced a probability of each biological condition category for each reachcode centroid. This was used to model an overall estimate of the percentage of Pennsylvania streams in each biological condition category with error estimation, which was the goal of this study.

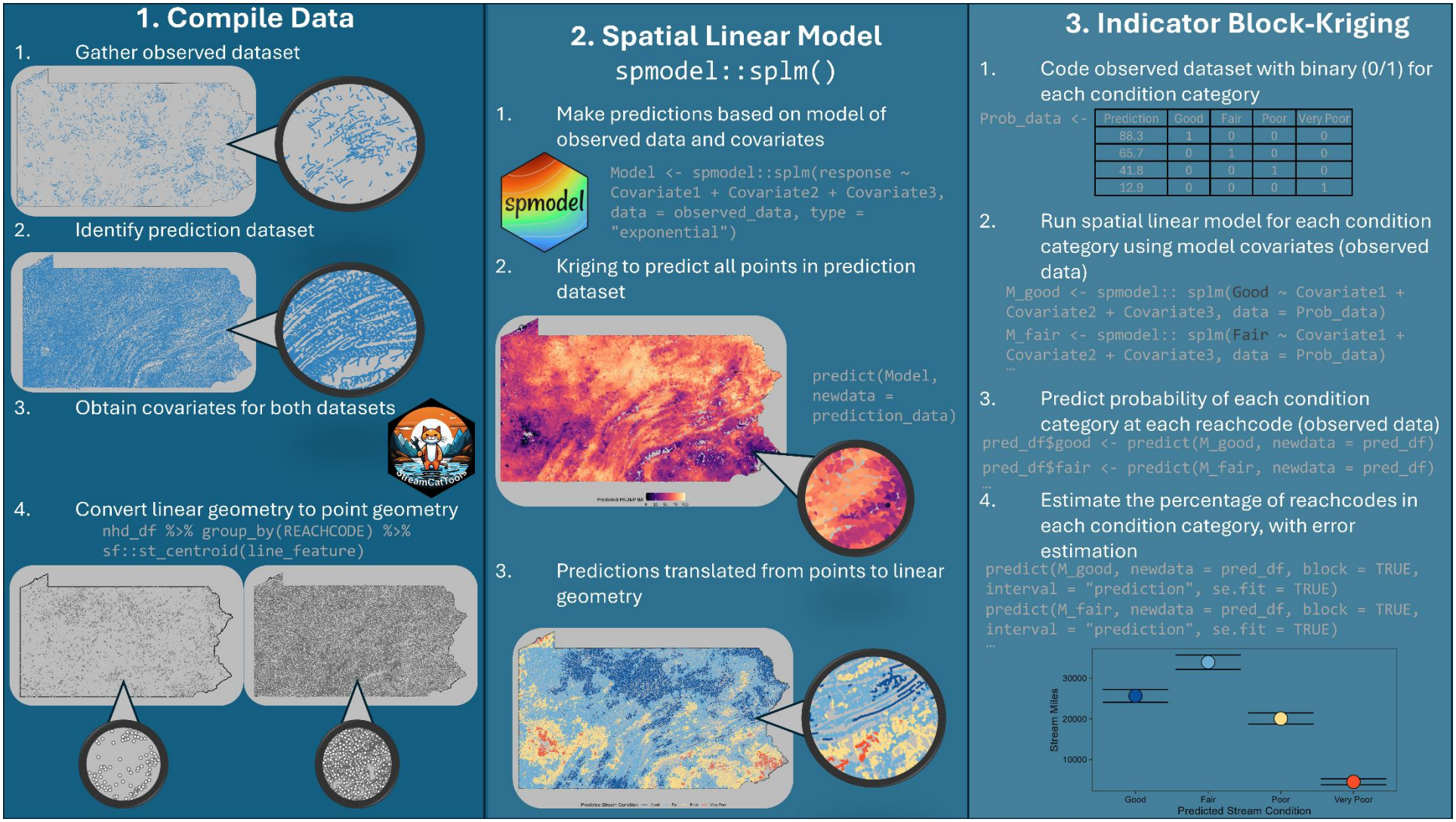

Figure 1. Flow chart of necessary steps for spatial linear modeling and indicator block-Kriging. Steps are organized from left-to-right and include details on R packages, R code, and specific descriptions of each step.

Spatial Datasets

Benthic Macroinvertebrates

DEP's largest macroinvertebrate community dataset, collected using the *Wadeable Riffle-Run Stream Macroinvertebrate Data Collection Protocol* (Shull 2017) and quantified using the Index of Biotic Integrity (IBI) from the *Wadeable Freestone Riffle-Run Stream Macroinvertebrate Assessment Method* (Shull 2023), was selected to build and validate the biological condition models. This macroinvertebrate IBI is measured on a 0-to-100-point scale where 0 represents the worst biological conditions and 100 represents the best biological conditions. The DEP collection protocol represents wadeable, perennial, freestone (i.e., relatively high gradient with little to no karst geology) watersheds up to 1,000 mi^2. In total, these samples represent the largest proportion of streams in Pennsylvania, which is roughly 84,300 streams miles out of the total 85,470 streams miles (96.6%) present based on the 1:24,000 High Resolution NHD.

To accurately reflect the target population that macroinvertebrate samples in freestone streams represents, a limit of 1,000 mi^2 was applied to both DEP samples and stream segments. Ephemeral/Intermittent streams represent a very small proportion of the NHD dataset in PA, so no reductions of the NHD dataset or samples were needed to facilitate accurate model performance and population condition estimates. Ultimately, the results of this model-based approach would represent perennial streams with established macroinvertebrate communities indigenous to freestone streams; it would not represent ephemeral/intermittent headwater streams or large rivers with drainages greater than 1,000 mi^2. Limestone and low-gradient streams <1,000 mi^2 were not removed from the NHD dataset due to the difficulty of determining the extent of these reaches without site-specific data. There is some potential for error since these stream types with distinct monitoring protocols and assessment methods were included; however, this error is expected to be small due to the relatively limited distribution of limestone and low-gradient streams. We also expect responses in these stream types to be consistent with freestone streams, so predictions of condition in these unique stream types is consistent with the goals of model-based prediction of stream condition.

The macroinvertebrate IBI dataset was then separated into two distinct time periods - one from 2006 to 2015 and one from 2016 to 2025 - allowing for separate models to be built and tested on the separate time periods as well as potential trend detection. The DEP macroinvertebrate dataset was attributed to the stream segment level of the NHD using the common identifier (COMID) attribute. One or more COMIDs comprise each reachcode, which are continuous stretches of streams or rivers with breaks resulting from confluences with tributaries or lakes/ponds. For each time period, macroinvertebrate IBI results were aggregated to the NHD High Resolution reachcode scale by first taking the mean of IBI scores for each COMID, then when multiple COMIDs were sampled on the same reachcode, the mean IBI score of those COMIDs was calculated. This reachcode-level aggregation results in a consistent schema where 1) when COMIDs were sampled with variable intensity, each COMID was given equal weight in the calculation of the overall IBI score for the reachcode and 2) reachcodes represented stream reaches with consistent representativeness due to the lack of influence of intervening tributaries, with the exception of point sources which may result in localized influences.

Each dataset was further separated into a calibration and validation dataset by randomly removing 20% of the samples to create the validation portion. The calibration datasets were used to build the models, and the validation datasets were used to compare observed versus predicted results. This calibration/validation approach is a common way to assess predictive performance of a model for unobserved data (James et al. 2013, Kuhn and Silge 2022). In order to build the framework for modeling and output for predicted values (i.e., IBI score by reachcode), the linear features of each reachcode were transformed into points by using the st_centroid() function in the sf package (Pebesma 2018) to get the centroid of each reachcode less than 1,000 mi^2 in Pennsylvania (n = 102,905). This was done to meet the point data requirements of the spmodel package (Dumelle et al. 2023b) and maintained the spatial integrity of the NHD dataset by using the centroid of each reachcode.

Model Covariates

Several watershed scale land cover percentages were used as covariates in model development. The National Land Cover Dataset (NLCD) was extracted for each reachcode (Dewitz 2023). This land cover dataset was used because it could be calculated for every NHD High Resolution (1:24,000) reachcode and it has been shown that macroinvertebrate communities are directly responsive to these variables (Chalfant 2012, Shull 2017).[1] The land cover types calculated, expressed as a percentage of each watershed, were:

- percent forest (i.e., total forest = Deciduous Forest + Mixed Deciduous/Evergreen Forest + Evergreen Forest),
- percent agriculture (i.e., total agriculture = Pasture/Hay + Row Crop),
- percent developed (i.e., total developed = Developed, Open Space + Developed, Low Intensity + Developed, Medium Intensity + Developed, High Intensity).

Land cover data were acquired for each reachcode by spatially relating each NHD High Resolution (1:24,000) reachcode to the StreamCat database (at 1:100,000 scale) (Hill et al. 2016). The relation was established by using a spatial join where each NHD High Resolution (1:24,000) reachcode was related to the corresponding NHD 1:100,000 StreamCat catchment using the “Have their center in” Spatial Join function in ArcPro. Relating the NHD 1:100,000 land cover data to the NHD High Resolution (1:24,000) dataset created some errors in the smaller watersheds, but this error was expected to be minor after preliminary auditing. The 2011 NLCD was used for the 2006 to 2015 dataset and the 2019 NLCD was used for 2016 to 2025 dataset. The final dataset resulted in a statewide wadeable freestone prediction dataset of 102,724 NHD High Resolution reachcodes representing the total target population of 84,300 miles with watershed land cover percentages available from StreamCat which were used for both model calibration and for prediction of biological condition at unmeasured reachcodes.

[1] One requirement of spatial linear models is no missing covariates in the dataset (i.e., every monitored and unmonitored location needs to have all covariates). Even though some parameters are commonly measured with biological samples, such as DO, temperature, pH, and conductivity, those data were not available for 100% of the sites. Rather than exclude sites, DEP chose to evaluate if the land cover variables were sufficient to predict macroinvertebrate IBI scores. More land cover variables were not explored because the land cover covariate calculation process and subsequent modeling is computationally expensive. If the selected land cover variables did not perform well, others may have been calculated and tested.

Spatial Linear Model Calibration and Validation

Spatial linear models were fit for each time period using the splm() function in the spmodel package, using the calibration IBI dataset. The watershed land cover percentages were used as covariates in each model. The best spatial covariance type was tested by running two models for each time period, using exponential and Gaussian spatial covariance functions. The model having the best fit (i.e., lowest AIC) was used to make site-specific predictions using Kriging and statewide predictions using block-Kriging (Cressie 1993, described in detail later).

The best fit models were checked for normality of standardized residuals using Q-Q plots to verify model assumptions of normality. Once spatial linear models were run, Kriging was performed using the predict() function to predict IBI score at all reachcode centroids in Pennsylvania (n = 102,274). Next, linear models were created for each time period to compare Kriging IBI score predictions of the calibration dataset against the observed IBI score at reachcodes that were extracted for the validation dataset. The calibration dataset did not include observed data for the validation reachcodes, so these predictions provided a measure of model accuracy for "out-of-sample" (i.e., unobserved) data. The relationship of observed IBI score from validation reachcodes versus predicted IBI scores from the calibration model needed to have a high correlation coefficient to fulfill this model validation step. Comparisons of observed versus predicted were completed as recommended by Piñeiro et al. (2008).

Once model diagnostics were completed, the calibration and validation datasets were combined to create one statewide (full) model for each time period. The modeling process was repeated as above, with 1) spatial linear models run for each time period with land cover covariates, 2) model normality checks, and 3) Kriging performed to predict IBI scores at all statewide wadeable freestone reachcodes. After model runs for each time period using the full datasets, the predictions of each reachcode score were joined to the NHD High Resolution flowline feature for segments with drainage areas < 1,000 mi^2 to create maps of the spatial linear model predictions.

Dataset Representativeness and Bias Testing

To confirm the validity of the models, each macroinvertebrate IBI observed dataset was tested for representativeness and bias by comparing sampled data to broader statewide patterns. Watershed size and land cover distribution were selected as the two most important variables for testing. This is an important final validation step because, if the datasets are biased, predicted condition estimates may not be accurate, or any observed trends may be erroneous.

Watershed size representativeness was tested by comparing the statewide wadeable freestone prediction dataset's watershed size distribution to each macroinvertebrate IBI observed dataset watershed size distribution. Macroinvertebrate IBI datasets needed to generally represent the statewide wadeable freestone prediction dataset distribution. Additionally, there should not be any major bias toward a particular watershed size or difference between either macroinvertebrate IBI observed dataset watershed size distribution. For example, if the 2006 to 2015 dataset represented a significantly higher portion of watersheds greater than 500 mi^2 than either the statewide wadeable freestone prediction dataset or the 2016 to 2025 dataset, then bias would be present, and the model results of the two periods could not be confidently compared for trend purposes.

Similar tests were conducted for each macroinvertebrate IBI dataset's land cover distribution. Land cover representativeness was tested by comparing the statewide wadeable freestone prediction dataset's percent forest, agriculture, and developed distributions to that of each macroinvertebrate IBI observed dataset. Both macroinvertebrate IBI datasets needed to generally represent the statewide wadeable freestone prediction dataset distribution with respect to land cover percentages. Additionally, neither of the macroinvertebrate IBI dataset's watershed land cover percentage distributions should have a major bias toward a particular type of land cover. For example, if the 2006 to 2015 dataset represented a significantly higher portion of developed watersheds than either the statewide wadeable freestone prediction dataset or the 2016 to 2025 dataset, then bias would likely be present, and the model results of the two time periods may not be confidently compared for water quality trend purposes.

Final Block-Kriging Model

The spatial linear models predicted IBI scores for each reachcode in each time period. It also allowed for a suitable way to validate the spatial linear model concept using the calibration and validation macroinvertebrate IBI datasets; however, a final modeling step was needed because spatial linear modeling does not provide valid error estimation. This final modeling step estimated the percentage of stream miles in each biological condition category, along with valid error estimation for each time period. The specific approach used in this case was called indicator block-Kriging (Cressie 1993). The results provide the statistical information needed to meet the requirements of the 106 Supplemental Grant.

To prepare the observed datasets for indicator block-Kriging, all samples were coded 0 or 1 for each biological condition category. Stream condition categories were created as follows: good = IBI scores greater than 75, fair = IBI scores between 75 and 50, poor = IBI scores between 50 and 25, and very poor = IBI scores less than 25. For example, if the mean IBI score of a reachcode was 64.9, then good = 0, fair = 1, poor = 0, and very poor = 0 for that reachcode. Four spatial linear models were run for each time period, with the observed binary good, fair, poor, and very poor response variables using the forest, agriculture, and developed covariates using the splm() function. Those four models for each time period were then used in indicator block-Kriging models to estimate stream condition (with error estimation) for all statewide wadeable freestone reachcodes using the predict() function in spmodel with block = TRUE, interval = "prediction", se.fit = TRUE arguments.

The resulting predictions represented point estimates with upper and lower 95th percentile standard errors for each stream condition category in each time period. Then, the point estimates were multiplied by the stream mileage in the prediction dataset to estimate the stream miles in each condition category. These models were considered successful if the margin of error of each biological condition category was less than 10%. This threshold comes from requirement that states, "State surveys aim to achieve 90% confidence ±10%...." (USEPA 2008).

RESULTS

Spatial Data

When the macroinvertebrate IBI datasets were separated into two time periods, the 2006 - 2015 macroinvertebrate IBI dataset represented 5,389 reachcodes and the 2016 - 2025 macroinvertebrate IBI dataset represented 3,829 reachcodes (Figure 2). Land cover data were not available for one reachcode for each time period. The 5,388 reachcodes in the 2006 – 2015 dataset had DEP IBI scores ranging from 0.5 - 100 with broad ranges of forested (0 - 100%), agriculture (0 - 90.6%), and developed (0 - 96.3%) land cover and ranged in drainage area from 0.01 – 985.8 mi^2. The 3,828 reachcodes in the 2016 - 2025 dataset had DEP IBI scores ranging from 2.3 - 99.9 with broad ranges of forested (0 - 100%), agriculture (0 - 92.7%), developed (0 - 99.5%) land cover and ranged in drainage area from 0.01 – 988.9 mi^2. Although these datasets were not randomly selected across Pennsylvania, each dataset was well distributed across the entire area of the state (Figure 2).

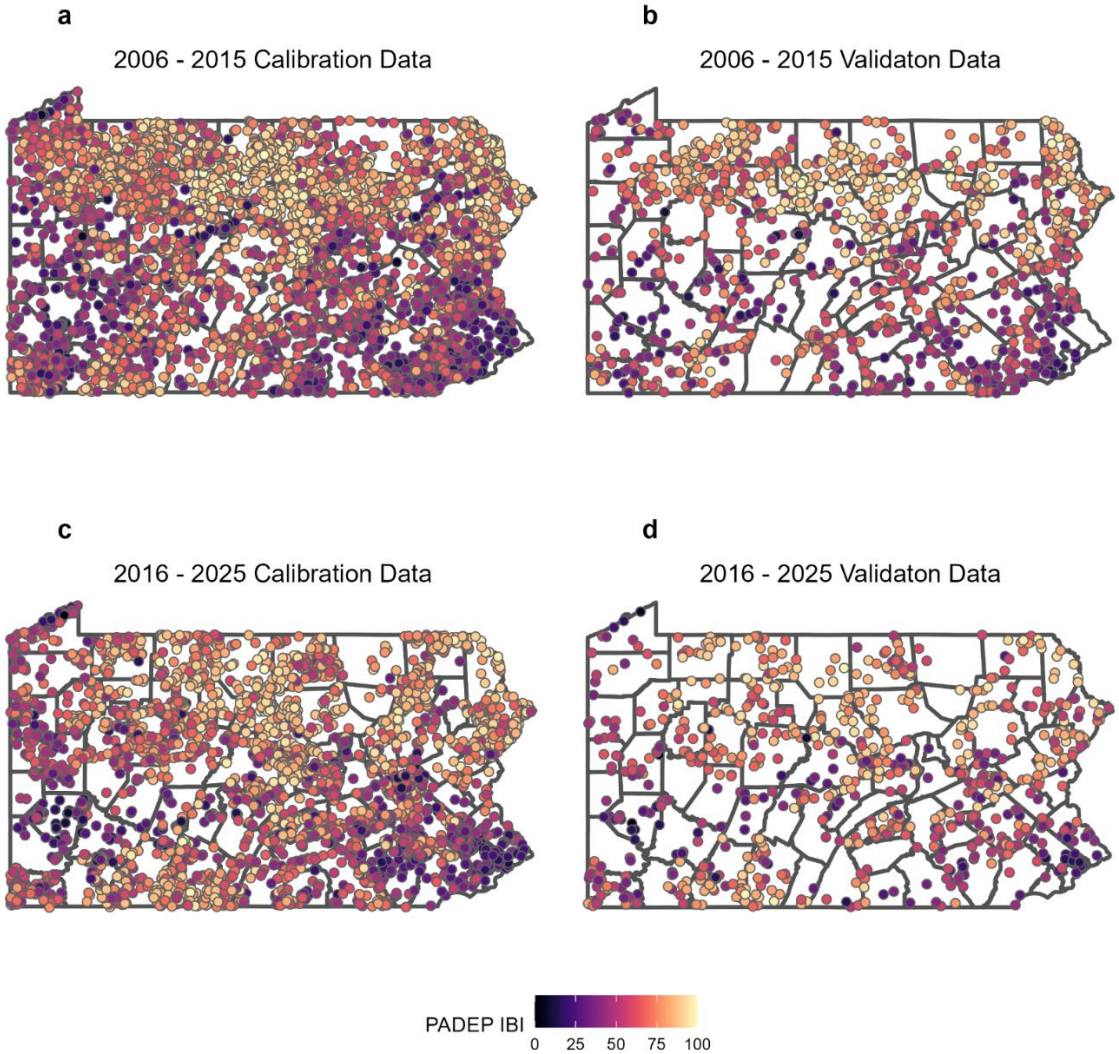

Figure 2. Mean DEP IBI scores for each available reachcode centroid in the 2006 – 2015 calibration (a), validation (b), and the 2016 – 2025 calibration (c) and validation (d) datasets.

Spatial Linear Model Calibration and Validation

Spatial linear models with land cover covariates were run for the 2006 - 2015 and 2016 - 2025 calibration macroinvertebrate IBI datasets. When 20% of the data were randomly removed for validation datasets, the calibration datasets had sample sizes of 4,311 and 3,064 for the respective time periods. Each model had the lowest AIC when the spatial covariance was set to exponential. Normality of spatial linear models standardized residuals were examined using Q-Q plots, and results suggested approximate-normality. The spatial linear model using 2006 - 2015 macroinvertebrate IBI dataset resulted in predicted DEP IBI scores for all 102,724 reachcodes ranging from -4.2 to 101.7. There were 39 (< 0.1%) predictions less than 0 and 16 (< 0.1%) predictions greater than 100. The spatial linear model using 2016 - 2025 macroinvertebrate IBI dataset resulted in predicted IBI scores for 102,724 reachcodes, ranging from -22.9 to 100.4. There were 24 (< 0.1%) predictions less than 0 and 5 (< 0.1%) predictions greater than 100. All predictions outside the 0-100 range of IBI score were truncated to 0 - 100 for subsequent visualization and analysis (Figure 3).

The validation step involved comparing relationships between observed macroinvertebrate IBI scores in the validation datasets and IBI scores predicted by the models using the calibration data. Across both models, predictions were highly correlated with observations (Pearson $R > 0.75$, $P < 0.001$) and the slope of predicted versus observed scores was very close to the 1:1 line (Figure 4). The validation step confirmed consistency between model predictions using calibration data and observed validation data. Therefore, calibration and validation datasets were combined, and spatial linear models were run for each time period using land cover covariates. The predictions of IBI score for each reachcode were joined to the NHD High Resolution flowline feature for segments with drainage areas < 1,000 mi^2 to create maps of the spatial linear model predictions. The predictions for both time periods aligned well spatially with land cover and known patterns of biological impairment in Pennsylvania (Figure 5).

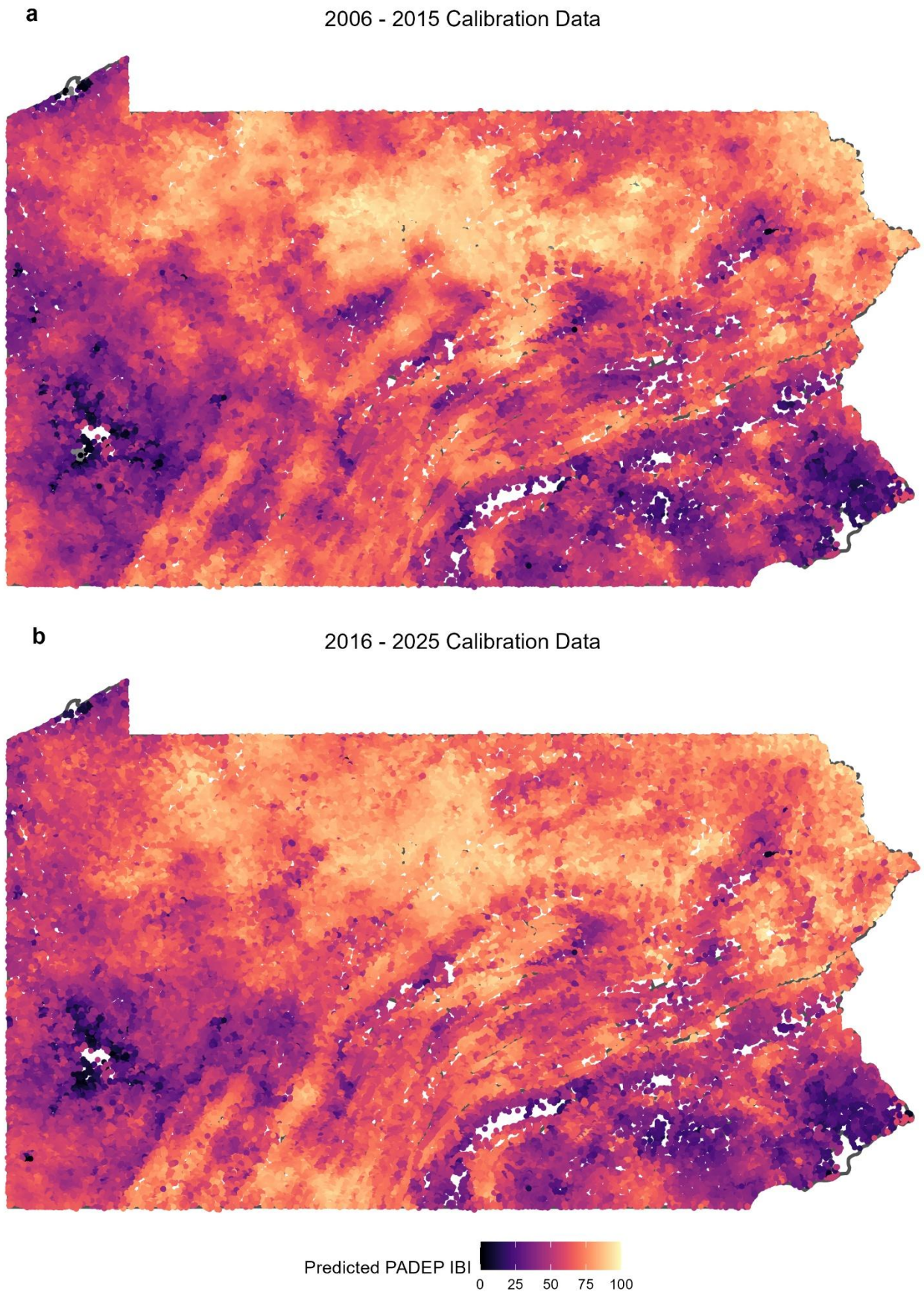

Figure 3. Predicted DEP IBI scores using 2006 - 2015 (top; a) and 2016 - 2025 (bottom; a) calibration models. Predicted values are shown as points for each reachcode centroid.

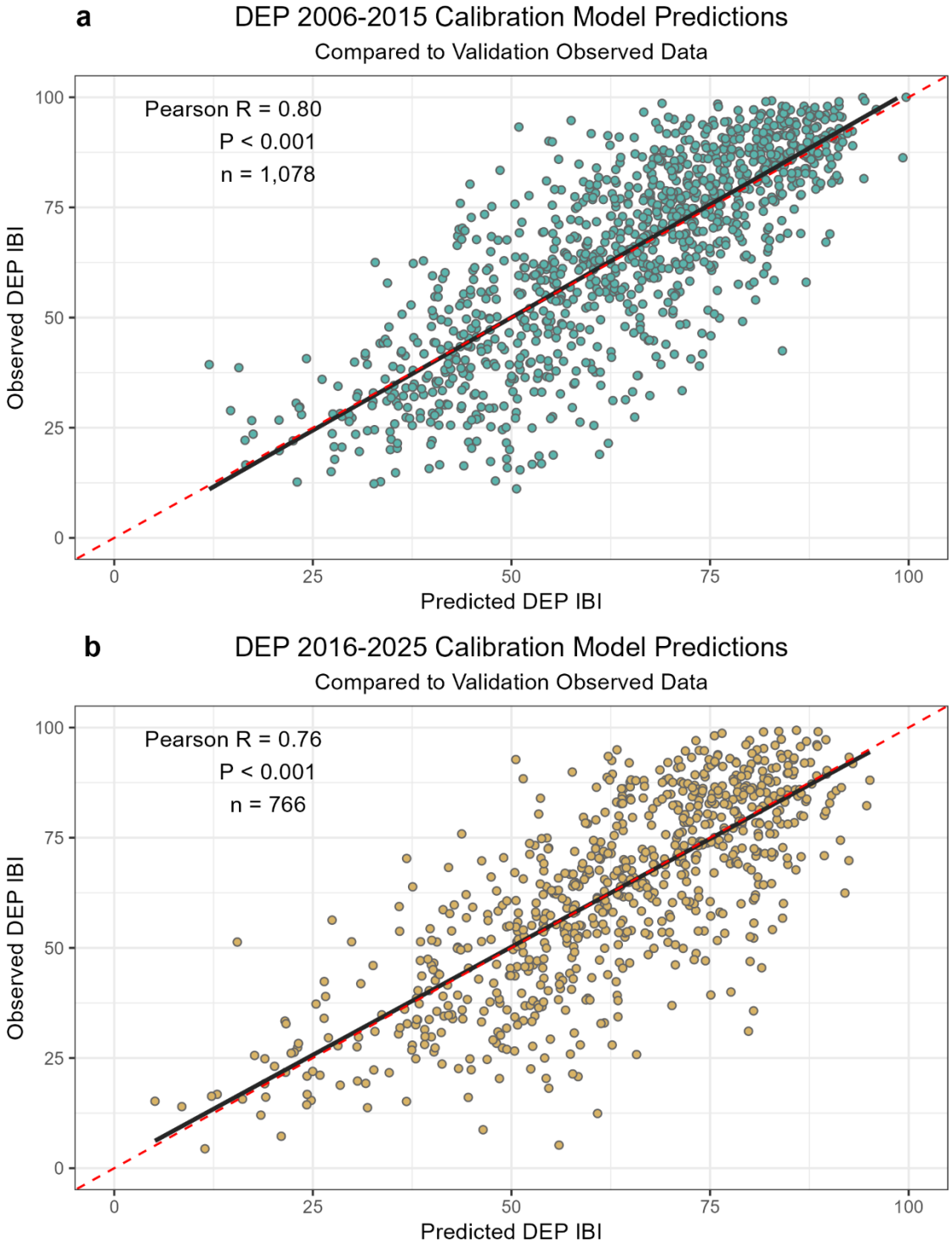

Figure 4. Relationship of observed DEP IBI (vertical axis) scores and predicted DEP IBI (horizontal axis) using 2006 - 2015 (top; a) and 2016 - 2025 (bottom; b) models. The predictions were generated using calibration models and observed data were from validation datasets. The dashed red line is the 1:1 line and the solid black line is the linear best fit line.

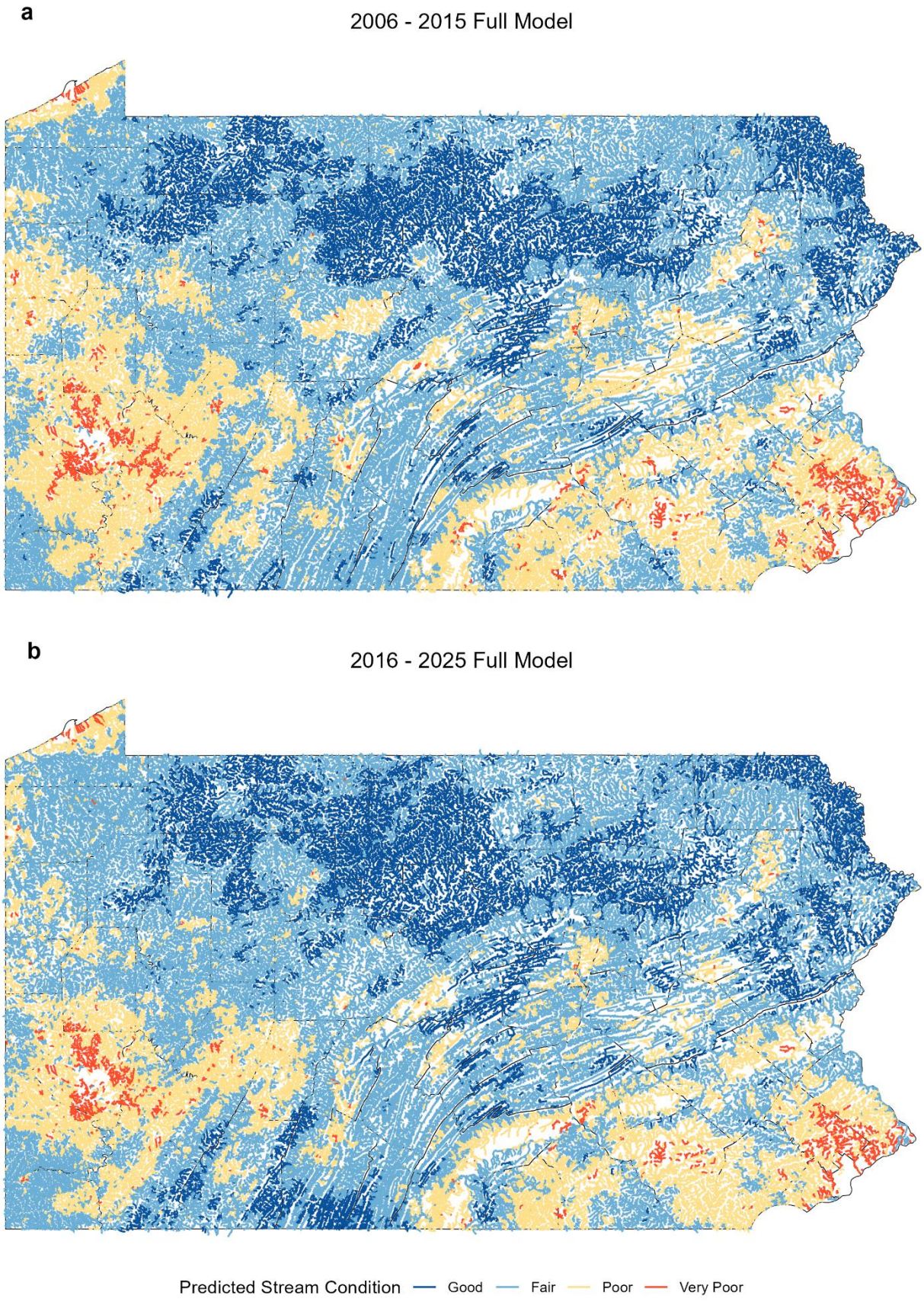

Figure 5. Predicted stream condition using combined calibration and validation datasets for 2006 - 2015 (top; a) and 2016 - 2025 (bottom; b) time periods. Spatial linear models with land cover covariates were used to make predictions.

Dataset Representativeness and Bias Testing

Both the 2006 to 2015 and 2016 to 2025 macroinvertebrate IBI observed datasets were compared to the statewide prediction dataset with respect to watershed size distribution (Figure 6). Results indicate that both macroinvertebrate datasets are most representative of watersheds less than 50 mi^2, which is consistent with the statewide prediction dataset's distribution, where approximately 90% of reachcodes represent watersheds less than 50 mi^2. The main difference between the statewide prediction dataset and both macroinvertebrate IBI observed datasets was that the statewide prediction dataset had a higher frequency of very small watersheds (approximately 0.1 to 2 mi^2), whereas both macroinvertebrate IBI datasets had the highest frequencies of watersheds between 2 to 5 mi^2. When compared together, the 2006 to 2015 dataset and 2016 to 2025 dataset were almost identical in watershed size distribution; however, very minor differences included a higher frequency of smaller watersheds (approximately 1 to 2 mi^2), and a higher frequency of the larger watersheds (greater than 500 mi^2) in the 2016 to 2025 dataset compared to the 2006 to 2015 dataset. It is also important to note that no drainage size groups were overrepresented within either of the macroinvertebrate IBI datasets.

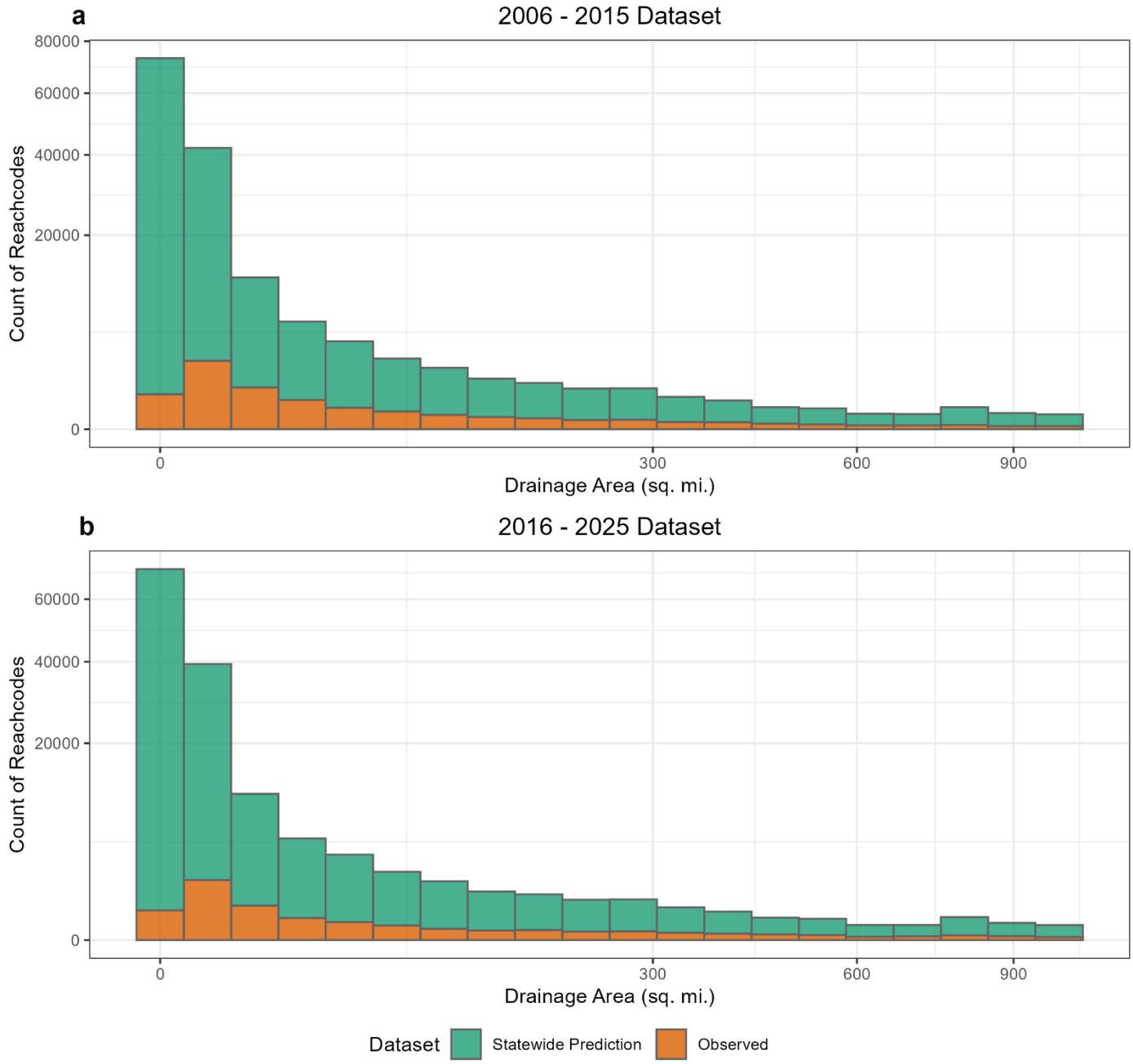

Figure 6. The frequency distributions for the watershed size of the statewide wadeable freestone prediction dataset compared to the 2006-2015 macroinvertebrate observed dataset (a), and the frequency distribution for the watershed size of the statewide wadeable freestone prediction dataset compared to the 2016-2025 macroinvertebrate observed dataset (b).

With respect to watershed land cover percentages, frequency distributions of both macroinvertebrate IBI datasets compared to their respective NLCD statewide wadeable freestone prediction datasets followed similar patterns for land cover categories (Figure 7). One slight difference was that both macroinvertebrate IBI datasets somewhat overrepresented watersheds with forested land cover between 65% to 90% compared to the statewide dataset. Agriculture and developed land covers in both macroinvertebrate IBI observed datasets followed almost identical patterns as the statewide NLCD

datasets. When the two macroinvertebrate IBI observed datasets were compared to each other, all three land cover categories were very similar, suggesting little to no bias between time periods.

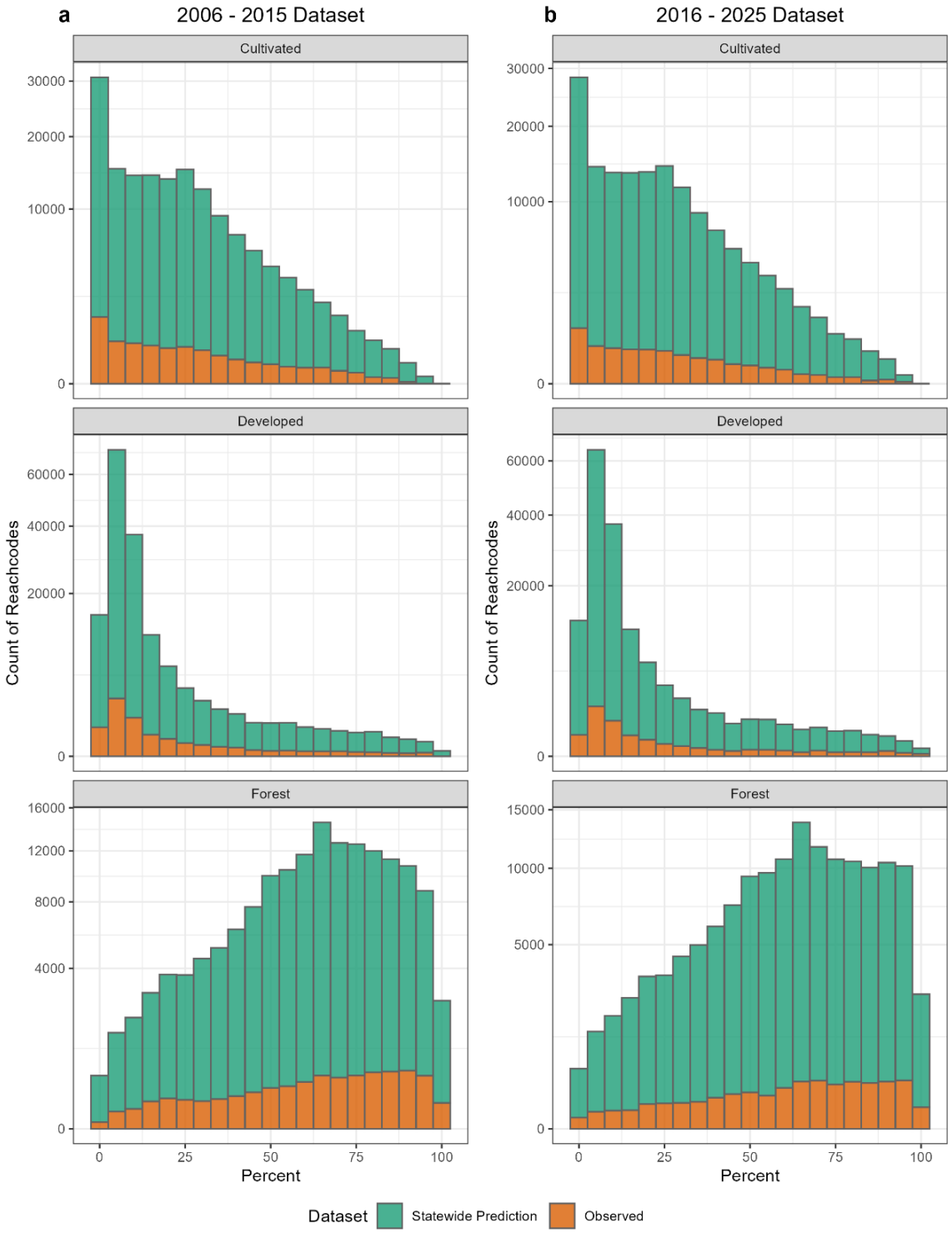

Figure 7. The frequency distributions for the 2011 NLCD forest, agriculture, and developed in the statewide prediction dataset compared to the same distributions of land cover in the 2006-2015 macroinvertebrate IBI observed dataset (a), and the frequency distributions for the 2019 NLCD forest, agriculture, and developed in the statewide prediction dataset compared to the same distributions of land cover in the 2006-2015 macroinvertebrate IBI observed dataset (b).

Final Block-Kriging Model

When respective calibration and validation datasets were combined and indicator block-Kriging models were applied, an almost identical percentage of streams were predicted to be in good condition in 2006-2015 and 2016-2025 (30.7% and 30.4%, respectively). The major difference between the models was fewer reachcodes were predicted to be in fair condition and more reachcodes were predicted to be in poor condition in 2006-2015 compared to the 2016-2025 period (Table 1; Figure 8). When estimates were expressed as miles, by multiplying the predicted stream condition percentage by the total stream mileage in the prediction dataset (84,322 miles), 25,872 miles were estimated to be in good condition, 30,053 were fair, 24,103 were poor, and 4,671 were very poor in 2006-2015. In 2016-2025, 25,661 miles were estimated to be in good condition, 33,951 were fair, 20,083 were poor, and 4,541 were very poor (Table 1). The 95% standard errors ranged from 0.77 to 2.08%, which were well below the 10% threshold needed to meet the recommendations of the 106 Supplemental Grant.

Table 1. Predicted stream condition in the 2006-2015 and 2016-2025 time periods with upper and lower 95th standard errors.

Time Period	Predicted Stream Condition	Estimate (%)	Lower 95th (%)	Upper 95th (%)	95% SE	Estimate (miles)	Lower 95th (miles)	Upper 95th (miles)
2006-2015	Good	30.68	28.69	32.68	1.99	25872	24190	27553
2006-2015	Fair	35.64	33.8	37.48	1.84	30052	28498	31607
2006-2015	Poor	28.58	26.91	30.26	1.67	24103	22692	25514
2006-2015	Very Poor	5.54	4.77	6.31	0.77	4671	4024	5318
2016-2025	Good	30.43	28.58	32.29	1.86	25660	24096	27224
2016-2025	Fair	40.26	38.19	42.34	2.08	33951	32199	35702
2016-2025	Poor	23.82	22.18	25.45	1.64	20082	18701	21463
2016-2025	Very Poor	5.39	4.51	6.26	0.87	4541	3803	5278

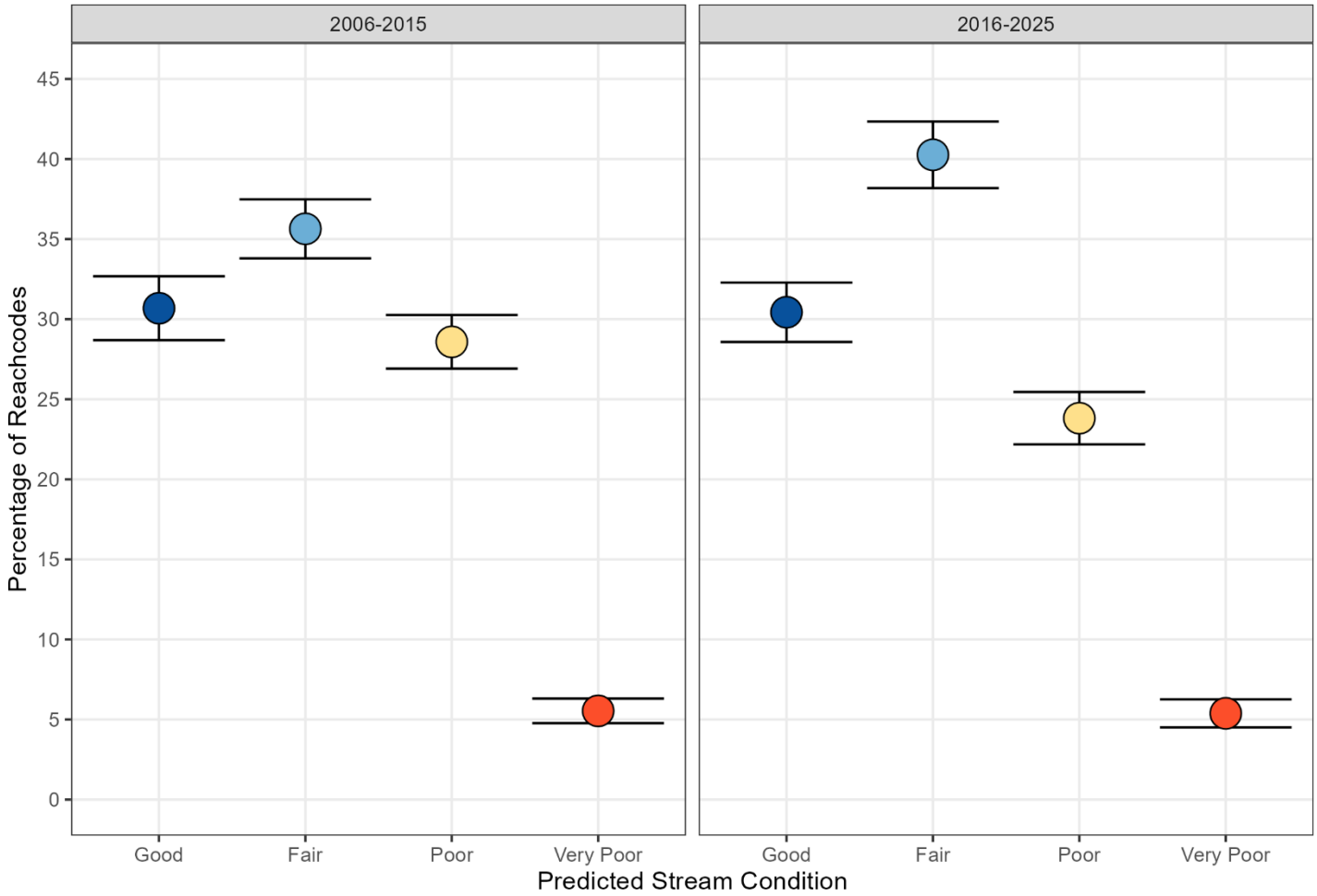

Figure 8. Prediction of percentage of reachcodes in each stream condition category in 2006-2015 (left pane) and 2016-2025 (right pane). Error bars represent 95th percentile standard errors.

DISCUSSION

Although the observed datasets spanned two distinct 10-year time periods, the spatial linear and indicator block-Kriging model predictions were consistent with known spatial patterns of biological conditions in Pennsylvania. The indicator block-Kriging models predicted nearly identical percentage of reachcodes (~30.5%) to be in good condition in both time periods. However, the percentage of fair reachcodes increased by nearly 5% and poor reachcodes decreased by nearly 5% from 2006-2015 to 2016-2025. The National Assessment for River and Streams report in 2019 produced similar macroinvertebrate community conditions results where 30% of the state was classified good, 30% was classified fair, and 40% was classified as poor (roughly 10% margin of error for each classification, USEPA 2023). This coherence with the National Assessment for River and Streams report helps to lend support to our model-based approach. The increased percentage of fair streams and decreased percentage of poor streams in the 2016 – 2025 period could indicate potential restoration and/or incremental improvement through time. This is somewhat consistent with expectations considering DEP has seen significant improvement in many of the poor biological conditions related to acid mine drainage over the last two decades. It is also interesting to note that other recent attempts to estimate biological condition in the larger Chesapeake Bay Watershed (of which Pennsylvania has a relatively

large watershed percentage) showed a likely improvement of roughly 6% during similar periods (Buchannan et al. 2023). For an in-depth understanding of temporal changes, reachcodes that switched conditions between time periods (indicating either improvement or degradation) could be examined to investigate the contributing factors. Sampling locations, restoration efforts, point sources, land cover changes, or other factors related to changes in water quality may become apparent as contributing factors. This type of in-depth examination could be beneficial to informing DEP's water quality management program. As time progresses and more data are collected, water quality trends could be paired with biological trends to provide a comprehensive picture of changing environmental conditions across the state.

It should be noted that these results are only predictions on the macroinvertebrate IBI score, which can be different than an Aquatic Life Use assessment decision on the same waterbody for several reasons. One reason is due to the extra assessment questions added to the method beyond IBI score impairment threshold of 50. If any one of the four assessment questions results in an impaired decision, the waterbody will be impaired, even if the IBI score is greater than 50. Additionally, the impairment threshold for IBI score in Special Protection streams is 63 (Shull 2023). Further, these predictions may be different than the overall Aquatic Life Use assessment decision found in DEP's Integrated Water Quality Monitoring and Assessment Report (DEP 2024). One of the main reasons for this is that DEP conducts Aquatic Life Use assessments using as many types of data and assessment methods as possible, and assessment methods are independently applicable when making a final assessment decision for a surface water. For example, if the macroinvertebrate IBI was greater than 50 and passed all the assessment screening questions, the surface water could still be impaired by physicochemical data that showed an exceedance of the total aluminum criteria found in Pennsylvania Code Title 25, Section 93.7.

Overall, observed macroinvertebrate IBI datasets for both time periods were considered representative and minimally biased when compared to statewide prediction datasets. The difference between watershed size representation in the statewide prediction dataset and the macroinvertebrate IBI datasets was expected considering the collection protocol was designed to target perennial streams with established macroinvertebrate communities and not ephemeral/intermittent headwater streams. Even with this underrepresentation, it should be appropriate to extrapolate condition to the very small streams when the land covers are similar, which in many cases during the preliminary auditing, they were. When watershed size was compared between the two time periods of macroinvertebrate IBI datasets, the differences were surprisingly minimal. Discrepancies between forest land cover representation in the statewide NLCD statewide prediction datasets and the observed macroinvertebrate IBI datasets were somewhat noticeable, with higher observed dataset representation in watersheds with higher forested percentages. This could present some bias toward predicting slightly higher biological conditions in the overall results, but the forested land cover distribution in both time periods of observed datasets were nearly identical, so this was not a concern from the trend perspective. When considered as a whole, the similarities far outweighed the differences, suggesting that predicted biological condition and trends between the time periods were not likely related to bias between the observed and prediction datasets.

Our conclusion is that the indicator block-Kriging approach in the spmodel package is an easily implementable and powerful statistical approach for classifying stream condition in unsampled reaches, as long as necessary precautions are undertaken to ensure the data are representative of the broader population of Pennsylvania streams being modeled. The potential utility for using this methodology to understand changes in stream condition at broad scales could be used to inform the DEP's water management program in numerous ways. Importantly, this work has demonstrated a statistically valid and novel approach to meet the requirements of the 106 Supplemental Grant funding.

LITERATURE CITED

Buchanan, C., R. D. Jepsen, and M. E. Mallonee. 2023. Stream Biological Health in the Chesapeake Bay Watershed. ICPRB Report ICP23-1. Report Prepared for the Chesapeake Bay Program Stream Health Workgroup.

Chalfant, B. 2012. A benthic index of biotic integrity for wadeable freestone streams in Pennsylvania. Pennsylvania Department of Environmental Protection, Harrisburg, Pennsylvania.

Cressie, N. 1993. Statistics for Spatial Data. New York. Wiley.

(DEP) Pennsylvania Department of Environmental Protection. 2024. 2024 Pennsylvania Integrated Water Quality Report. Pennsylvania Department of Environmental Protection, Harrisburg, Pennsylvania. Available online at https://storymaps.arcgis.com/stories/7af67824d6924b88b544dbad302ebc4f

Dewitz, J. 2023. National Land cover Database (NLCD) 2021 Products [Data set]. U.S. Geological Survey. Available online at https://doi.org/10.5066/P9JZ7AO3

Dumelle, M., M. Higham, J. M. Ver Hoef, A. R. Olsen, and L. Madsen. 2022. A comparison of design-based and model-based approaches for finite population spatial sampling and inference. Methods in Ecology and Evolution, 13(9), 2018-2029.

Dumelle M, Kincaid T, Olsen AR, Weber M 2023a. spsurvey: Spatial Sampling Design and Analysis in R. *Journal of Statistical Software*, 105(3), 1–29.

Dumelle M, Higham M, Ver Hoef JM 2023b. spmodel: Spatial statistical modeling and prediction in R. PLoS ONE 18(3).

Hill, Ryan A., Marc H. Weber, Scott G. Leibowitz, Anthony R. Olsen, and Darren J. Thornbrugh. 2016. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. Journal of the American Water Resources Association (JAWRA) 52:120-128.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning: with applications in R. New York: Springer.

Kuhn, M. and J. Silge. 2022. Tidy modeling with R: A framework for modeling in the tidyverse. O'Reilly Media, Inc.

Nahlik, A. M., S. G. Paulsen, M. Dumelle, S. Holdsworth, S. Lehmann, N. Tulve, S.Paul, and H. C. Frey. 2025. National Aquatic Resource Surveys (NARS): the foundation for long-term aquatic monitoring data across the United States. Environmental Monitoring and Assessment, 197(12), 1291.

Pebesma E. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal*, 10(1), 439–446.

Piñeiro, G., S. Perelman, J. P. Guerschman, J.M. Paruelo. 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed? Ecological Modelling, 216 (3–4) 316-322.

Shull, D. R. 2017. A benthic macroinvertebrate multimetric index for large semiwadeable rivers technical report. Pennsylvania Department of Environmental Protection, Harrisburg, Pennsylvania.

Shull, D. R. (editor). 2017. Wadeable riffle-run stream macroinvertebrate data collection protocol. Chapter 3.1 *in* M. J. Lookenbill and E. Arnold (editors). Water quality monitoring protocols for surface waters. Pennsylvania Department of Environmental Protection, Harrisburg, Pennsylvania.

Shull, D. R., and M. Pulket (editors). 2023. General Assessment Considerations. Chapter 1.2, pages 15–21 *in* D. R. Shull, and R. Whiteash (editors). Water quality assessment methodology for surface waters. Pennsylvania Department of Environmental Protection, Harrisburg, Pennsylvania.

United States Environmental Protection Agency. 2008. Guidelines for the Award of Monitoring Initiative Funds Under Section 106 Grants to States, Interstate Agencies, and Tribes. 73 Fed. Reg. 41071.

United States Environmental Protection Agency. 2023. National Rivers and Streams Assessment 2018–2019: Pennsylvania Results Unpublished Report. Office of Water, Office of Research and Development, Washington, DC.

Ver Hoef, J. M. 2008. Spatial methods for plot-based sampling of wildlife populations. Environmental and Ecological Statistics, 15(1), 3-13.

Yang, X., and W. Jin. 2010. GIS-based spatial regression and prediction of water quality in river networks: A case study in Iowa. Journal of Environmental Management. 91(10), 1943-1951.

Yang, X., Q. Liu, X. Lou, and Z. Zheng. 2017. Spatial Regression and Prediction of Water Quality in a Watershed with Complex Pollution Sources. Scientific Reports. 7, 8318.

Zaresefat, M., R. Derakhshani, and J. Griffioen. 2024. Empirical Bayesian Kriging, a Robust Method for Spatial Data Interpolation of a Large Groundwater Quality Dataset from the Western Netherlands. Water. 16(18), 2581.

Zhang, Z., Q. Du. 2019. A Bayesian Kriging Regression Method to Estimate Air Temperature Using Remote Sensing Data. Remote Sensing. 11(7), 767.

Zimmerman, D. L., and J. M. Ver Hoef. 2024. Spatial linear models for environmental data. Chapman and Hall/CRC.