Spora: A Journal of Biomathematics

Volume 7 Article 8

2021

Empirical Fitting of Periodically Repeating Environmental Data

Pavel Bělík
Augsburg University, belik@augsburg.edu

Andrew Hotchkiss
Augsburg University, andy.hotchkiss@alumni.augsburg.edu

Brandon Perez
Augsburg University, brandon.cruz.perez@gmail.com

John Zobitz
Augsburg University, zobitz@augsburg.edu

Empirical Fitting of Periodically Repeating Environmental Data

Cover Page Footnote

The idea for this study arose from a student project taught in *Calculus I* by Co-authors Zobitz and Bělík. Funding for Co-authors Hotchkiss and Perez was provided by Augsburg's University Undergraduate Research and Graduate Opportunity. Co-author Zobitz acknowledges B. S. Chelton for helpful comments on this manuscript.

Intercollegiate Biomathematics Alliance

SPORA
A Journal of Biomathematics

ISSN 2473-3067 (Print)
ISSN 2473-5493 (Online)

Empirical Fitting of Periodically Repeating Environmental Data

Pavel Bělík[1], Andrew Hotchkiss[1], Brandon Perez[1], John Zobitz[1,*]

*Correspondence:
Augsburg University,
2211 Riverside Ave.,
Minneapolis, MN 55454, USA
zobitz@augsburg.edu

Abstract

We extend and generalize an approach to conduct fitting models of periodically repeating data. Our method first detrends the data from a baseline function and then fits the data to a periodic (trigonometric, polynomial, or piecewise linear) function. The polynomial and piecewise linear functions are developed from assumptions of continuity and differentiability across each time period. We apply this approach to different datasets in the environmental sciences in addition to a synthetic dataset. Overall the polynomial and piecewise linear approaches developed here performed as good (or better) compared to the trigonometric approach when evaluated using statistical measures (R^2 or the AIC). These results were consistent when the number of measurements decreased (through random removal of data). Future applications of the fitting method could account for higher-order terms in the polynomial function or refinements to the estimation of parameters in the piecewise linear function.

Keywords: periodic timeseries, polynomial functions, piecewise linear functions, net carbon uptake, evapotranspiration

1 Introduction

Biological systems can exhibit periodic behavior in their physiological patterns (such as internal circadian rhythms [23, 41]) or due to phenological or seasonal changes [5]. In many circumstances, collection of nearly-continuous datasets demonstrate processes that can be classified as periodic and span a range of disciplines from ecology [13], atmospheric science [39], or neuroscience [8]. These datasets are then used for modeling and analysis activities [17, 19, 21, 31, 37].

Consider for example Figure 1, a timeseries of net carbon uptake for a forest in Colorado. Negative values imply a net carbon loss from the ecosystem to the atmosphere (which signifies conversion of atmospheric carbon by plants). Figure 1 shows the long-term (baseline) trend is for this ecosystem to absorb carbon. However there is a distinct annual pattern where the net carbon is increasing through the wintertime periods, but decreases (almost monotonically) during the summer months.

A way to mathematically model this periodic behavior is with combinations of trigonometric functions of a known period [7]. This approach is an example of empirical curve fitting, which parameterizes measured data $(t_i,\ y_i)$ with a function $y = F(t, \vec{\alpha})$, where $\vec{\alpha}$ is a set of unknown parameters. Usually a least squares criterion is applied that minimizes the difference between the function $y = f(x)$ and measured data [10]. When the data exhibit

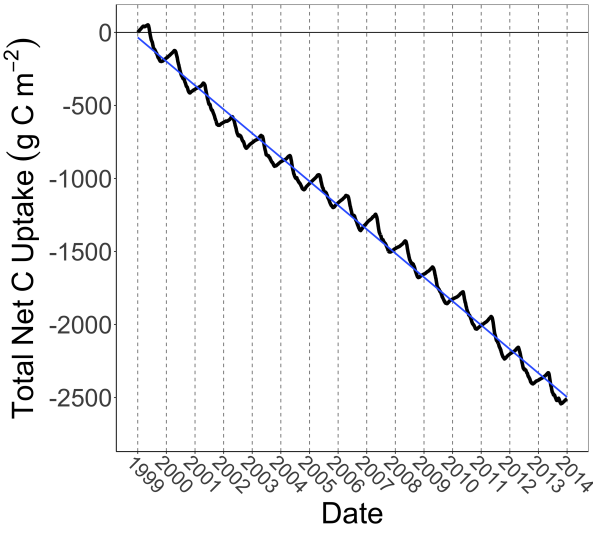

Figure 1: Timeseries of cumulative net carbon uptake for a high elevation coniferous forest in Colorado. This periodically repeating timeseries shows the ecosystem is slowly accumulating carbon, with an annual period of decrease due to actively absorbing carbon during the summer months. A long-term linear trend is plotted with the data (blue line).

[1]Department of Mathematics, Statistics, & Computer Science, Augsburg University, Minneapolis, MN

a long term trend (such as a linear decrease in Figure 1) one approach is to first "detrend" the data by determining a function to represent the long-term trend [15, 42] and then conducting the empirical curve fit to the residual between $\vec{y}$ and the long-term trend.

Following detrending of data, empirical curve fitting procedures of periodic data can be done through a variety of approaches, often using trigonometric functions as a starting point in the analysis. Edwards [9] uses an approach that relies on transforming the data to a unit circle and examining the probability distribution from that transformation. David and Bliss [7] generalizes the trigonometric approach with Fourier analysis. Huang et al. [14] generalizes orthogonal least squares regression to determine the phase shift in a data, especially applied to circadian rhythms. Additionally, because the vast majority of these datasets have time as a predictor, principles of timeseries analysis can be applied [16].

For this manuscript we develop and evaluate alternative approaches for empirical fitting of periodically repeating data without using trigonometric functions. Our approach utilizes polynomial functions to first detrend the long-term cycle, although many other types of functions could conceivably be used. We would expect some degree of overfitting as the degree of the baseline polynomial increases, so we evaluate the minimum degree of the polynomial applied for detrending the long-term trend. For the periodic components, our approach specifies the degree of smoothness at the endpoints of the period (for Figure 1 this would mean the start and end of the annual cycles be both continuous and differentiable). Along with trigonometric functions, we develop a polynomial and a piecewise linear approach for fitting the periodic components. We evaluate our approaches to a range of nearly continuous datasets found in the ecological literature.

2 Methods

2.1 Datasets utilized

We used a variety of datasets that span a range of different environmental observational data, along with a synthetic dataset for demonstration of the fitting procedure.

2.1.1 Cumulative Net Ecosystem Carbon Exchange

The data set utilized here is the net carbon uptake from a high-elevation coniferous forest in Colorado (Figure 1; Monson et al. [24]). Measurement of the net carbon uptake is derived from mass conservation and micrometeorological and biophysical principles, which can be represented as an example of the Fundamental Theorem of Calculus [43]. The net carbon uptake is influenced by whole ecosystem respiration (which causes the cumulative carbon uptake to increase) and gross primary productivity (which causes the net carbon uptake to decrease). Units of the net carbon uptake are g C m^{-2}. A broad network of sites (FLUXNET) measure the net carbon uptake and other associated measurements (`www.fluxdata.org`). For this study we use daily measurements of the net carbon uptake. We will refer to this dataset as "Total Net C Uptake".

2.1.2 Mauna Loa carbon dioxide

The Mauna Loa carbon dioxide dataset is a long-term record of directly measured CO_2 mole fraction (from 1958 to the present day) from the Mauna Loa observatory in Hawaii [18, 39]. For this analysis we use the monthly mean value of CO_2 (units parts per million or ppm) provided by NOAA (`www.esrl.noaa.gov/gmd/ccgg/trends/`). We will refer to this dataset as "CO_2". Figure 4 includes the CO_2 data used in this study.

2.1.3 Evapotranspiration data

The third data set is evapotranspiration (denoted as ET, units mm H_2O m^{-2}), which is the combined sum of water lost to the atmosphere through evaporation (from soils, water runoff, or plant interception) and plant transpiration. Over a landscape the rate of transpiration can be inferred from satellite remote sensing products collected by NASA MODIS Terra and Aqua satellites [6, 11, 22, 25, 25], provided on an eight-day timescale [26, 35, 36]. Values of ET at a specific location are derived from a model that takes into consideration the surface vegetation, daily meteorological variables and surface reflectance. We accessed the ET data product through the A$\rho\rho$EEARS web service (`https://lpdaacsvc.cr.usgs.gov/appeears/`, A$\rho\rho$EEARS Team [2]), for a deciduous forest located in Australia (Site AU-Lox, 34.4707° S and 140.6551° E, `DOI:10.18140/FLX/1440247`) from 2012 to 2018. The ET data product was filtered using the highest quality assurance flags (e.g., when there are no clouds present) which narrows down the number of days in which reflectance data were utilized. For this site we would expect the pattern in ET to be periodic due to the seasonality in leaf-on and leaf-off at this site. Visual inspection of the data (Figure 5) does not seem to indicate a long-term increasing or decreasing trend in the data, in contrast to Total Net C uptake (Figure 1). We will refer to this dataset as "ET". Figure 5 includes the ET data used in this study.

2.1.4 Synthetic Data

Finally, we formulated a synthetic dataset spanning ten years with an annual period. This annual cycle was com-

posed of three piecewise linear functions with breakpoints 20% and 60% through the year; the slope on the first and third segments was set to $m = 5$ (similar to the piecewise linear function in Figure 2). Superimposed on this annual cycle was a long-term trend as decreasing linear function. Finally the output (y) values were randomly perturbed with normally distributed random noise with mean zero and standard deviation 0.1. We will refer to this dataset as "Synthetic". Figure 6 includes the Synthetic data used in this study.

2.2 Fitting procedure

The fitting procedure takes periodic data $\vec{d} = \{t_i, y_i\}$ with period ρ to approximate the function $y = F(t, \vec{\alpha})$, where $\vec{\alpha}$ is a vector of parameters determined by the fitting routine. We assume that F includes both a periodic component $P(t, \vec{\alpha}_P)$ and a baseline component $B(t, \vec{\alpha}_B)$, which will be a polynomial (function, so we assume that $F(t, \vec{\alpha}) = B(t, \vec{\alpha}_B) + P(t, \vec{\alpha}_P)$ and $\vec{\alpha}$ is the union of $\vec{\alpha}_B$ and $\vec{\alpha}_P$. Unless specified otherwise, we applied ordinary least squares linear regression for the fitting procedure.

The fitting procedure first removes the long-term trend from the data by fitting the baseline component $B(t, \vec{\alpha}_B)$ and computing the residual between the data and $B(t, \vec{\alpha}_B)$. The residual data are used to parameterize $P(t, \vec{\alpha}_P)$, which are assumed to be periodic. We then define $\tau = t \bmod \rho$ be a scaled variable between zero and unity. In this case we only need to parameterize the function for the scaled variable $P(\tau, \vec{\alpha}_P)$. After parameterizing $P(\tau, \vec{\alpha}_P)$, we then rescale P in terms of the variable t. We consider three different models for $P(\tau, \vec{\alpha}_P)$, as described in the next three sections.

2.2.1 Trigonometric model

The function $P(\tau, \vec{\alpha}_P)$ is approximated with trigonometric functions, so

$$P(\tau, \vec{\alpha}_P) = a_0 + a_1 \sin(2\pi\tau) + a_2 \cos(2\pi\tau). \qquad (1)$$

Equation (1) is a standard model based on David and Bliss [7]. The Trigonometric model has three parameters (a_0, a_1, and a_2) determined by our fitting procedure.

2.2.2 Polynomial model

The function $P(\tau, \vec{\alpha}_P)$ is approximated by a polynomial function so

$$P(\tau, \vec{\beta}) = \sum_{i=0}^{n} \beta_i \tau^i \cdot (1-\tau)^{n-i}$$

with $n \geq 1$. We write the function in this way to ensure periodicity at $\tau = 0$ and $\tau = 1$.

In order to ensure this polynomial function is continuous and differentiable on the interval $0 \leq \tau \leq 1$ we apply the constraints $P(0, \vec{\beta}) = P(1, \vec{\beta}) = c$ and $P'(0, \vec{\beta}) = P'(1, \vec{\beta}) = m$. These constraints determine conditions on the parameters β_i for any degree polynomial:

$$\begin{aligned} P(0, \vec{\beta}) &= \beta_0 = c, \\ P(1, \vec{\beta}) &= \beta_n = c, \\ P'(0, \vec{\beta}) &= -n\beta_0 + \beta_1 = m, \\ P'(1, \vec{\beta}) &= -\beta_{n-1} + n\beta_n = m. \end{aligned}$$

A cubic polynomial is the minimum degree polynomial for $P(\tau, \vec{\alpha}_P)$ that satisfies these constraints, but in this instance $P(\tau, \vec{\alpha}_P)$ is symmetric about $\tau = 0.5$, which might limit the applicability of this method. As a result, we fix $n = 4$ for this study; after simplification we have

$$\begin{aligned} P(\tau, \vec{\alpha}_P) = m\tau(1-\tau)(1-2\tau) \\ + c(1 - 6\tau^2 + 12\tau^3 - 6\tau^4) \\ + \beta_2 \tau^2 (1-\tau^2). \end{aligned} \qquad (2)$$

Consequently, even though the Polynomial model is a quartic (degree 4) model, it also has three parameters (m, c, and β_2) determined by our fitting procedure.

2.2.3 Piecewise Linear model

A third model we consider for $P(\tau, \vec{\alpha_P})$ is a piecewise linear function. We still assume continuity conditions as with the Polynomial model:

$$P(\tau, \vec{\alpha}_P) = \begin{cases} m\tau + c & 0 \leq \tau < b_1 \\ m\left(\frac{b_2 - b_1 - 1}{b_2 - b_1}\right)(\tau - b_1) & \\ \qquad + mb_1 + c & b_1 \leq \tau < b_2 \\ m\,(\tau - 1) + c & b_2 \leq \tau \leq 1 \end{cases} \qquad (3)$$

In Equation (3) the parameters b_1 and b_2 are considered the breakpoints where $P(\tau)$ changes slope. We first determine the location of these breakpoints with segmented linear regression [27, 28, 30]. While algorithms for segmented linear regression can determine breakpoints and slopes on each segment, to ensure differentiability we need the slope (m) to be equal on the first and third segments. Also, to keep the number of parameters comparable to the other two models, we only consider the case with two breakpoints. More breakpoints can easily be considered for more other datasets. Once these breakpoints are determined Equation (3) is linear with parameters m and c. As a result, the Piecewise Linear model has four parameters (m, c, b_1, and b_2) determined by our fitting procedure.

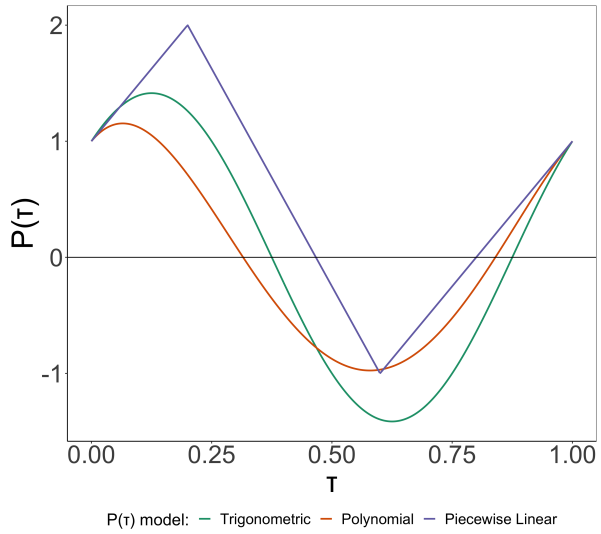

Figure 2: Representative plot of the Trigonometric (Equation (1)), Polynomial (Equation (2)) and Piecewise Linear (Equation (3)) models. Values of $a_1 = a_2 = c = 1$ and $m = 5$, $b_1 = 0.2$, $b_2 = 0.6$.

A representative plot for the different $P(\tau)$ models is shown in Figure 2.

2.3 Software utilized

We utilized R [32], RStudio [34], and the tidyverse (Wickham et al. [40]; `http://tidyverse.tidyverse.org`) for data processing and analysis. Specific R packages included tidyverse for data processing and visualization. Breakpoints b_1 and b_2 in Equation (3) were determined with the segmented package [28]. All the code used to generate results and figures can be found at `https://github.com/jmzobitz/periodicFitting`.

2.4 Model evaluation approaches

We will evaluate our model results by how well the different models (Trigonometric, Periodic, and Piecewise Linear) reproduce the periodic data, taking into consideration the baseline function used for the fit. For model evaluation we computed summary statistics (such as R^2 values) and Akaike's Information Criterion (denoted as AIC, Akaike [1]) to select the best approximating fitted model for each dataset. The R^2 is interpreted as the proportion of the measured residual variance accounted for by the model; R^2 values close to unity suggests the modeled variance closely approximates the measured variance. The AIC is computed as AIC$= 2p - 2\ln(L)$, where p is the number of parameters fit to the data and L the model likelihood (in this case it is the modeled variance).

Because our measurements (and model fits) vary in time periodically, a second model comparison examines the differences in the variability in the *pattern* of modeled to measured data. This is called the normalized centered root mean square difference, denoted here as σ_n [38]. The value of σ_n is the standard deviation of the difference between two residuals: the measurements and the fitted values. The residual for both is the difference between an individual value and the overall mean (in this case the measurements or the fitted values).

Beyond summary statistics we also compared which of the three approaches (Trigonometric, Polynomial, or Piecewise Linear) better approximated the periodic trends as data become more sparse. First, we fixed the baseline function $B(t)$ with the lowest AIC in Table 1. We also investigated the effect of the fitting procedure (Trigonometric, Polynomial, and Piecewise Linear) when the number of measurements in a dataset decreases. For each dataset we randomly sampled a percentage of the original data and then re-applied our fitting procedures. This process was repeated 500 times and the ensemble average was computed to create a distribution of the AIC (for each fit) as a function of the proportion of data removed.

3 Results

We computed a Trigonometric, Polynomial, and Piecewise Linear fit for each of the datasets in Section 2.1 using baseline models ranging from a constant function to a fourth degree polynomial. Each dataset was fitted twelve different ways (four baseline functions each for the Trigonometric, Polynomial, and Piecewise Linear fits) and summary statistics were computed.

Figures 3–6 show the results of applying the fitting procedures described in Section 2.2. Overall the fitting procedure (Trigonometric, Polynomial, or Piecewise Linear) was able to visually reproduce the trends in the given timeseries. In most cases (with the exception of Figure 5) non-constant polynomial functions for $B(t)$ produced a better visual representation for the data. The detrended timeseries for the Piecewise Linear model fit had $c = 1.1$, $m = 4.7$, $b_1 = 0.21$, and $b_2 = 0.59$. The fitted values were close to the original values of $c = 1.0$, $m = 5$, $b_1 = 0.2$, $b_2 = 0.6$ (Figure 2). As expected for the Synthetic dataset, the Piecewise Linear model had the best model-data fit (lowest σ_n, Table 1), although the Trigonometric and Polynomial models performed comparably well (Figure 6).

Table 1 reports summary statistics (R^2, AIC, and σ_n) for each model. For instances where the AIC is equal for a dataset (e.g., the Total Net C Uptake data) we chose the lower degree of $B(t)$. In contrast to the other datasets,

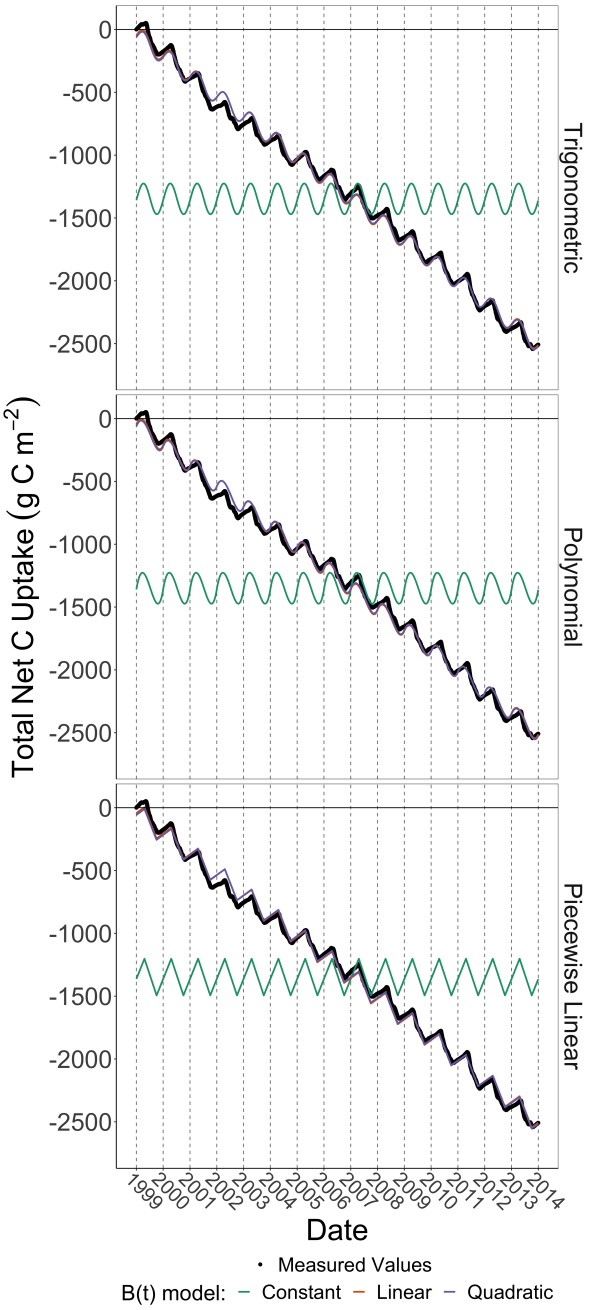

Figure 3: Model results for Total Net C Uptake using different baseline functions $B(t)$ (colored lines) along with measured values (black dots) by applying the fitting routine described in Section 2.2. We omitted baseline fits beyond quadratic functions due to the high similarity in results.

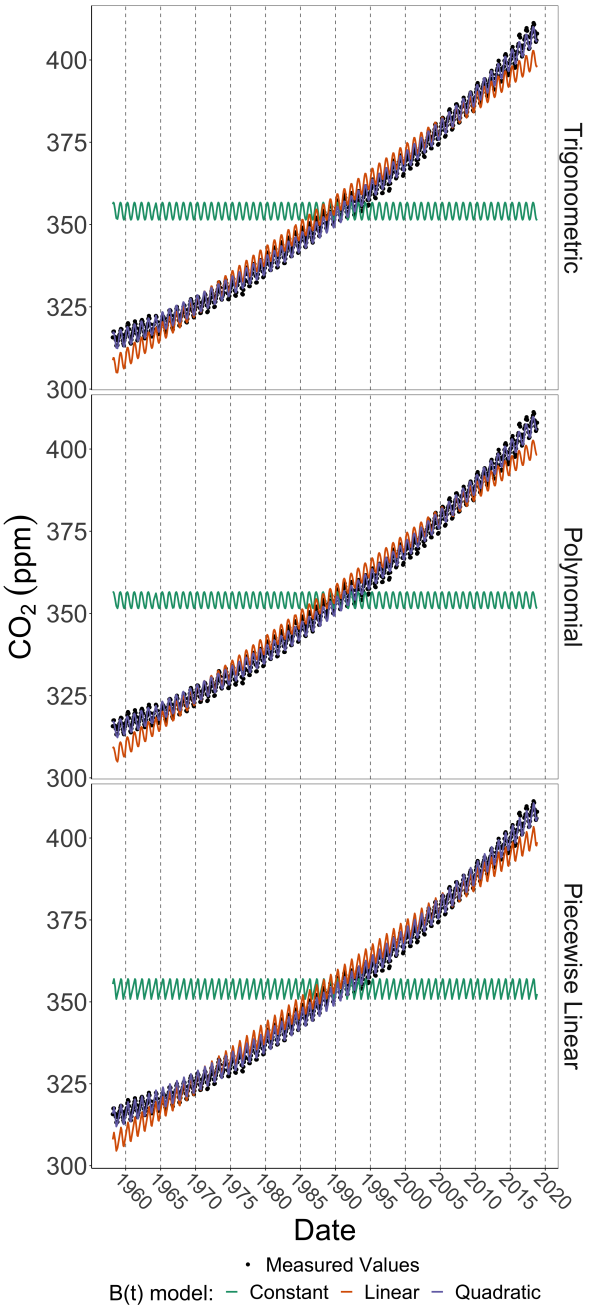

Figure 4: Model results for Mauna Loa CO_2 data using different baseline functions $B(t)$ (colored lines) along with measured values (black dots) by applying the fitting routine described in Section 2.2. We omitted baseline fits beyond quadratic functions due to the high similarity in results.

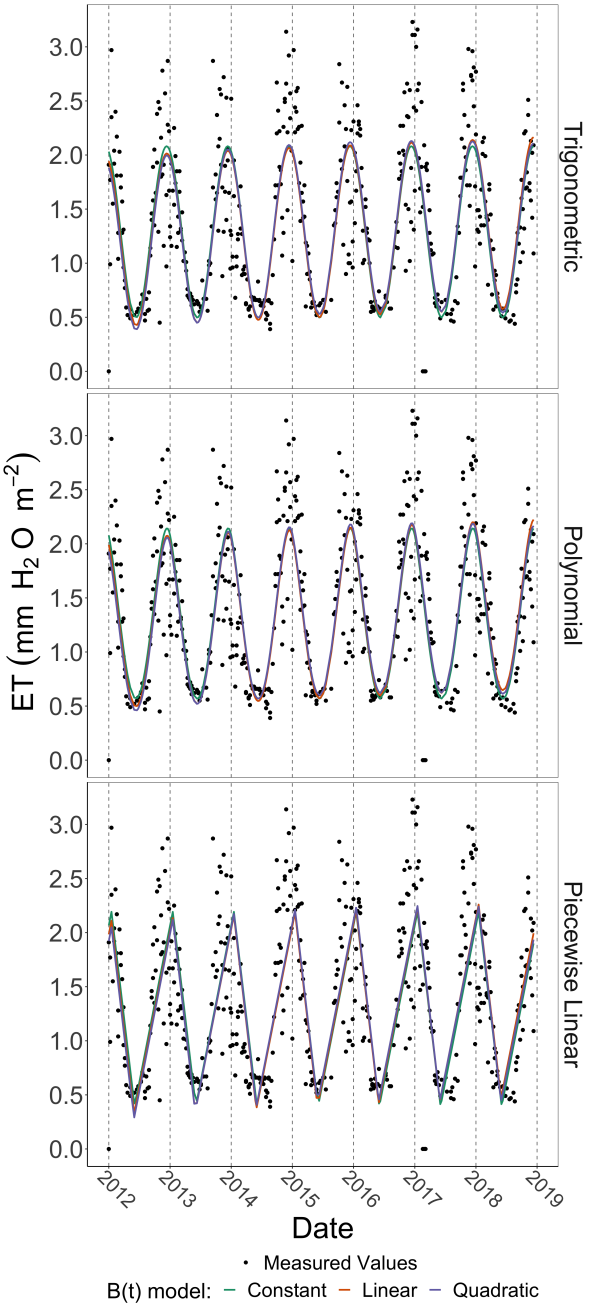

Figure 5: Model results for ET using different baseline functions $B(t)$ (colored lines) along with measured values (black dots) by applying the fitting routine described in Section 2.2. We omitted baseline fits beyond quadratic functions due to the high similarity in results.

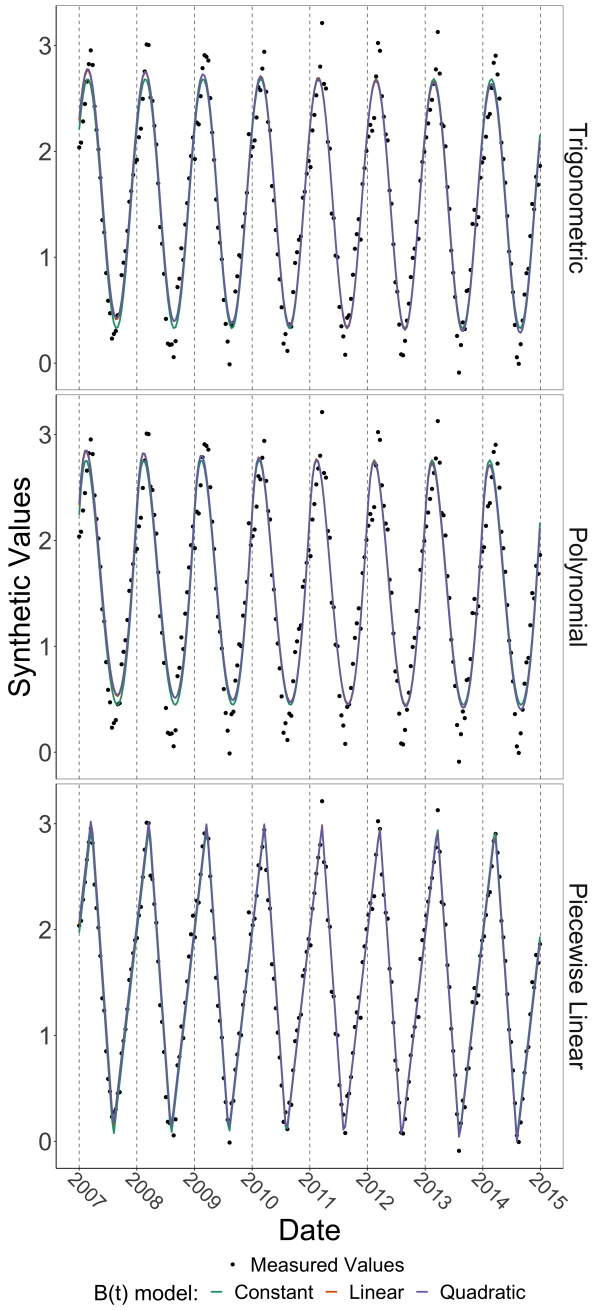

Figure 6: Model results for the Synthetic dataset using different baseline functions $B(t)$ (colored lines) along with measured values (black dots) by applying the fitting routine described in Section 2.2. We omitted baseline fits beyond quadratic functions due to the high similarity in results.

the AIC values for the Synthetic dataset were negative. This is because the likelihood L is very large, which in turn made the AIC (computed as $2p - 2\ln(L)$) a positive value.

The Total Net C Uptake and CO_2 datasets exhibit a strong periodic trend, so the R^2 for non-constant polynomials is already very high. For the Synthetic data, the Piecewise Linear model had the lowest σ_n, which is to be expected since this dataset was constructed from piecewise linear functions.

Figure 7 shows results when we re-evaluated each model with sparse datasets. For the CO_2 and Synthetic datasets the same model was preferred (Piecewise Linear for CO_2 data or Polynomial for Synthetic data) even as the gaps in the dataset increased. For the Total Net Carbon Uptake and ET datasets all methods performed comparatively well, with perhaps more variation in the computed AIC for the Piecewise Linear approach.

4 Discussion

This study presents alternative approaches to empirical model fitting of periodic datasets with polynomial functions or piecewise linear functions using differentiability and continuity constraints. On the whole, these approaches performed as well as fitting with trigonometric functions (Figures 3–6), with several cases of the fitted functions being indistinguishable from each other. The values in Table 1 show increasing the degree of the baseline function $B(t)$ does not significantly improve the overall data fit, with the AIC preferring a lower degree polynomial for $B(t)$. The baseline function $B(t)$ selected did not depend on the approach utilized.

The fits (Trigonometric, Polynomial, or Piecewise Linear) from the ET dataset, comparatively speaking, produced worse model fits compared to the other datasets. We attribute this difference to the high variability in this dataset, both shown in Figure 5 and with the largest value of σ_n in Table 1, which reports the normalized center root mean square difference for the fit with the lowest AIC for each fitting approach. One way to reduce the residual variance for the Trigonometric model is by including higher order terms in Equation 1 (e.g., $\sin(m\pi\tau)$, $\cos(m\pi\tau)$, where m is an integer). An analogous approach to the Periodic model would be to increase the degree of the polynomial n. Future work could investigate the reduction in the residual variance as additional terms are added for noisy datasets such as ET. We would not expect additional terms in the Trigonometric or Polynomial models to significantly improve results for datasets such as Total Net C Uptake or CO_2 because there already is a high representation of the fitted values to the measurements.

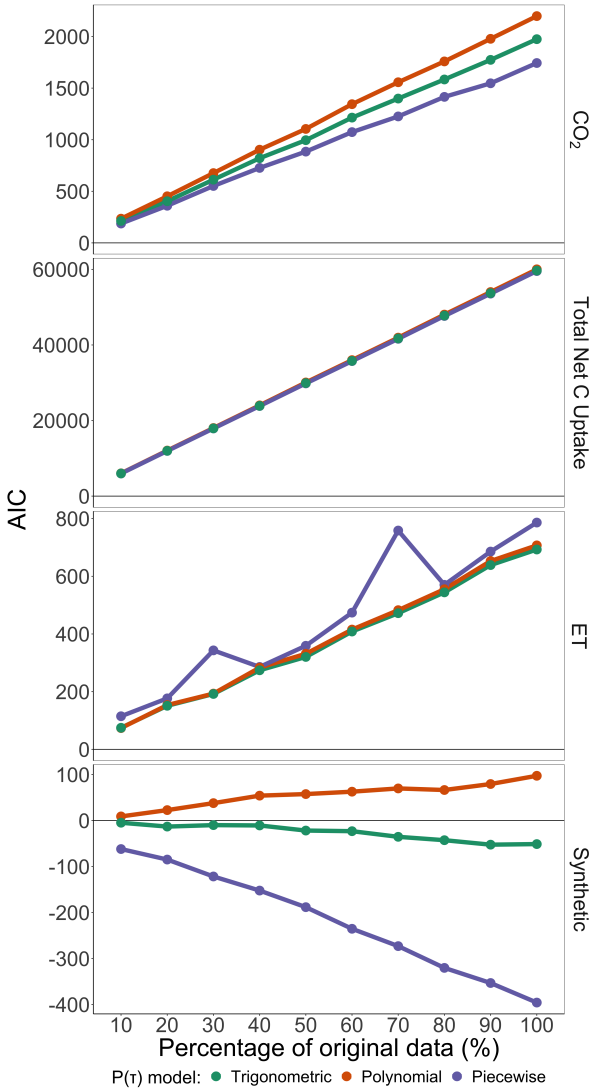

Figure 7: Ensemble median value of AIC values for the different fitting procedures by dataset with a proportion of data randomly removed before fitting. For each fitting procedure and data the baseline function $B(t)$ with the lowest AIC in Table 1 was utilized.

Table 1: Model summary results (including the R^2 and AIC (Akaike Information Criterion) organized by equation used to estimate the periodic component (either trigonometric, Equation (1); polynomial, Equation (2); or piecewise linear (3). For each fitting approach the degree of $B(t)$ that has the highest R^2 (blue color) or lowest AIC (red color) is highlighted. (In the case of ties the lowest degree for $B(t)$ is selected.) The last column σ_n reports normalized center root mean square difference for the fit with the lowest AIC (red cell in each row).

Statistic	R^2					AIC					σ_n
$B(t)$ degree → $P(t)$ function ↓	0	1	2	3	4	0	1	2	3	4	
Total Net C Uptake											
Trigonometric	0.01	1	1	1	1	94073.54	59864.8	59693.55	59695.55	59697.55	0.05
Polynomial	0.01	1	1	1	1	94070.42	60179.69	60017.71	60019.71	60021.71	0.05
Piecewise Linear	0.01	1	1	1	1	94079.5	59736.08	59560.84	59562.84	59564.84	0.05
CO_2											
Trigonometric	0	0.98	1	1	1	6840.63	3905.42	1975.47	1974.59	1976.59	0.03
Polynomial	0	0.98	1	1	1	6840.88	3923.46	2198.16	2198.15	2200.15	0.04
Piecewise Linear	0	0.98	1	1	1	6842.38	3893.1	1745.13	1743.46	1745.46	0.03
ET											
Trigonometric	0.56	0.56	0.57	0.57	0.57	692.83	695.68	695.5	697.5	699.5	0.66
Polynomial	0.55	0.55	0.55	0.55	0.55	707.14	709.96	710.08	712.08	714.08	0.67
Piecewise Linear	0.48	0.49	0.49	0.49	0.49	790	781.85	782.51	784.51	786.51	0.72
Synthetic											
Trigonometric	0.94	0.93	0.93	0.93	0.93	−51.13	−32.6	−30.09	−28.09	−26.09	0.26
Polynomial	0.89	0.89	0.89	0.89	0.89	97.35	110.48	112.81	114.81	116.81	0.33
Piecewise Linear	0.98	0.98	0.98	0.98	0.98	−395.67	−345.23	−342.35	−340.35	−338.35	0.15

The Polynomial approach can be generalized further by requiring additional constraints beyond continuity and differentiability. Just like higher-order harmonic terms can be added with the Trigonometric fitting approach, more general, higher-degree polynomials could be conceivably used in the Polynomial approach. One would then have to be careful to choose a model that doesn't introduce unwanted spurious oscillations within each period. For example, one idea is to use $\tau(1-\tau)^k - \tau^k(1-\tau)$ in place of $\tau(1-\tau)^2 - \tau^2(1-\tau) = \tau(1-\tau)(1-2\tau)$. This has a nice effect of spreading the max/min farther apart, which may be desirable for some datasets. In this case k needs to be a hyperparameter that is selected ahead of time. Similarly, the Piecewise Linear model could be extended for datasets with a unique maximum and a unique minimum in each period of the detrended data by using higher-degree polynomials on each interval where the data is approximately monotonic. Future work could extend this approach to integrate with other curve fitting approaches (e.g., non-parametric estimation, Hall et al. [12]; timeseries analysis, Kovács et al. [20], Wu et al. [42]). Additionally, to ensure $P(\tau)$ has biologically reasonable solutions, additional constraints could incorporate constraints on the parameters β_2 and m into the fitting routine.

The Piecewise Linear model assumes equal slopes for the intervals $0 \le \tau \le b_1$ and $b_2 \le \tau \le 1$ to ensure $P(\tau)$ is differentiable at $\tau = 1$. In segmented regression analysis, slopes on different intervals are not assumed to be equal [27, 29, 30]. While Equation (3) is nonlinear with respect to parameters b_1, b_2, m, and c, for computational simplicity we decided to determine the breakpoints first and then apply linear regression with indicator variables to determine m and c. We believe this approach is more comparable to the Trigonometric and Periodic approaches. Future investigations could contrast a nonlinear optimization method to the approach outlined here.

Figure 7 shows that no model seemed to suddenly outperform the others as data gaps increase. The AIC changes in Figure 7 because the log-likelihood is changing (the number of parameters p is fixed because the baseline polynomial $B(t)$ is fixed). As a result, increases in the AIC are associated with a decreasing log-likelihood (suggesting a better model-data fit for the measurements; CO_2, Total Net C Uptake, *ET* datasets). In contrast, for the Synthetic dataset the log-likelihood increased as the data became less sparse, causing the AIC to decrease. The log-likelihood in our case is essentially the residual sum of squares, so removing data will decrease the model-data residual. The Synthetic data already exhibit a high

degree of periodicity (Figure 6), so randomly thinning the dataset would remove more outliers than compared to the other datasets.

With any dataset invariably measurement gaps can occur—either due to sampling frequency or intermittent instrument outages. Measurements of the carbon uptake (Figure 1) can have gaps as high as 35% of the measurement period [3, 4]. When this occurs, gap-filling techniques (such as [33]) are one way to generate approximately correct values consistent with expected patterns. Through thinning of the dataset, the Polynomial approach presented here performed as good or better than trigonometric functions (Figure 7), and the constraints described could be generalized further depending on individual knowledge of the system at hand.

Further investigation could examine the length of the interval $b_1 \leq \tau \leq b_2$ (where b_1 and b_2 could be the optimum values on the periodic interval or the breakpoints) and its influence on the choice of a particular fitting approach (Trigonometric, Polynomial, or Piecewise Linear). Apart from the Piecewise Linear approach, as a rule of thumb we found that the Trigonometric approach works better if the difference $b_2 - b_1 \approx 0.5$, and the Polynomial approximation works better when $b_2 - b_1$ is greater than 0.5. When $b_2 - b_1$ is less than 0.5 no approach was better than the other. Additional numerical simulation with synthetic data (such as what we did with the AIC in Figure 7) could establish baseline metrics for each of the fitting procedures.

Acknowledgments

The idea for this study arose from a student project taught in *Calculus I* by Co-authors Zobitz and Bělík. Funding for Co-authors Hotchkiss and Perez was provided by Augsburg's University Undergraduate Research and Graduate Opportunity. Co-author Zobitz acknowledges B. S. Chelton for helpful comments on this manuscript.

Author Contributions

Co-authors Zobitz and Bělík conceived the ideas for the research project, contributed to the study design, and wrote the manuscript. Co-authors Hotchkiss and Perez conducted the research project by analyzing the datasets, evaluating the results, and contributed to the writing of the manuscript.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[2] A$\rho\rho$EEARS Team (2018). Application for Extracting and Exploring Analysis Ready Samples (A$\rho\rho$EEARS). Ver. 2.14.2. `https://lpdaacsvc.cr.usgs.gov/appeears`.

[3] Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., U, K. T. P., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K., and Wofsy, S. (2001). FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. *Bulletin of the American Meteorological Society*, 82(11):2415–2434.

[4] Baldocchi, D., Reichstein, M., Papale, D., Koteen, L., Vargas, R., Agarwal, D., and Cook, R. (2012). The role of trace gas flux networks in the biogeosciences. *Eos Trans. AGU*, 93(23):doi:10.1029/2012EO230001.

[5] Chuine, I. and Régnière, J. (2017). Process-Based Models of Phenology for Plants and Animals. *Annual Review of Ecology, Evolution, and Systematics*, 48(1):159–182.

[6] Cleugh, H. A., Leuning, R., Mu, Q., and Running, S. W. (2007). Regional evaporation estimates from flux tower and MODIS satellite data. *Remote Sensing of Environment*, 106(3):285–304.

[7] David, F. N. and Bliss, C. I. (1959). Periodic Regression in Biology and Climatology. *Biometrika*, 46(3/4):495.

[8] de Cheveigné, A. and Arzounian, D. (2018). Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. *NeuroImage*, 172:903–912.

[9] Edwards, J. H. (1961). The recognition and estimation of cyclic trends. *Annals of Human Genetics*, 25(1):83–87.

[10] Golub, G. H. and Van Loan, C. (2014). *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition.

[11] Guay, K. C., Beck, P. S. A., Berner, L. T., Goetz, S. J., Baccini, A., and Buermann, W. (2014). Vegetation productivity patterns at high northern latitudes: A multi-sensor satellite data assessment. *Global Change Biology*, 20(10):3147–3158.

[12] Hall, P., Reimann, J., and Rice, J. (2000). Nonparametric Estimation of a Periodic Function. *Biometrika*, 87(3):545–557.

[13] Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., and Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3):156–162.

[14] Huang, Y., Bowman, C., Walch, O., and Forger, D. (2019). Phase estimation from noisy data with gaps. In *2019 13th International Conference on Sampling Theory and Applications (SampTA)*, pages 1–4.

[15] Iler, A. M., Inouye, D. W., Schmidt, N. M., and Høye, T. T. (2017). Detrending phenological time series improves climate–phenology analyses and reveals evidence of plasticity. *Ecology*, 98(3):647–655.

[16] Jones, R. H. and Brelsford, W. M. (1967). Time series with periodic structure. *Biometrika*, 54(3-4):403–408.

[17] Jung, M., Verstraete, M., Gobron, N., Reichstein, M., Papale, D., Bondeau, A., Robustelli, M., and Pinty, B. (2008). Diagnostic assessment of European gross primary production. *Global Change Biology*, 14(10):2349–2364.

[18] Keeling, C. D., Bacastow, R. B., Bainbridge, A. E., Ekdahl, C. A., Guenther, P. R., Waterman, L. S., and Chin, J. F. S. (1976). Atmospheric carbon dioxide variations at Mauna Loa Observatory, Hawaii. *Tellus*, 28(6):538–551.

[19] Kemp, S., Scholze, M., Ziehn, T., and Kaminski, T. (2014). Limiting the parameter space in the Carbon Cycle Data Assimilation System (CCDAS). *Geosci. Model Dev.*, 7(4):1609–1619.

[20] Kovács, G., Zucker, S., and Mazeh, T. (2002). A box-fitting algorithm in the search for periodic transits. *Astronomy & Astrophysics*, 391(1):369–377.

[21] Kuppel, S., Peylin, P., Maignan, F., Chevallier, F., Kiely, G., Montagnani, L., and Cescatti, A. (2014). Model–data fusion across ecosystems: From multisite optimizations to global simulations. *Geosci. Model Dev.*, 7(6):2581–2597.

[22] Kussul, N., Skakun, S., Shelestov, A., Lavreniuk, M., Yailymov, B., and Kussul, O. (2015). Regional scale crop mapping using multi-temporal satellite imagery. In *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XL-7/W3, pages 45–52, Berlin, Germany.

[23] Levi, F. and Schibler, U. (2007). Circadian Rhythms: Mechanisms and Therapeutic Implications. *Annual Review of Pharmacology and Toxicology*, 47(1):593–628.

[24] Monson, R. K., Turnipseed, A. A., Sparks, J. P., Harley, P. C., Scott-Denton, L. E., Sparks, K., and Huxman, T. E. (2002). Carbon sequestration in a high-elevation, subalpine forest. *Global Change Biology*, 8:459–478.

[25] Mu, Q., Zhao, M., Kimball, J. S., McDowell, N. G., and Running, S. W. (2013a). A Remotely Sensed Global Terrestrial Drought Severity Index. *Bulletin of the American Meteorological Society*, 94(1):83–98.

[26] Mu, Q., Zhao, M., and Running, S. W. (2013b). MODIS Global Terrestrial Evapotranspiration (ET) Product (NASA MOD16A2/A3) Algorithm Theoretical Basis Document Collection 5. Technical report, NASA.

[27] Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22:3055–3071.

[28] Muggeo, V. M. (2008). Segmented: An R package to fit regression models with broken-line relationships. *R News*, 8(1):20–25.

[29] Muggeo, V. M. (2016). Testing with a nuisance parameter present only under the alternative: A score-based approach with application to segmented modelling. *J of Statistical Computation and Simulation*, 86:3059–3067.

[30] Muggeo, V. M. (2017). Interval estimation for the breakpoint in segmented regression: A smoothed score-based approach. *Australian & New Zealand Journal of Statistics*, 59:311–322.

[31] Pinty, B., Lavergne, T., Vossbeck, M., Kaminski, T., Aussedat, O., Giering, R., Gobron, N., Taberner, M., Verstraete, M. M., and Widlowski, J. L. (2007). Retrieving surface parameters for climate models from Moderate Resolution Imaging Spectroradiometer (MODIS)-Multiangle Imaging Spectroradiometer (MISR) albedo products. *J. Geophys. Res*, 112:D10116.

[32] R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[33] Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D.,

Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., and Valentini, a. R. (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Global Change Biology*, 11:1424–1439.

[34] RStudio Team (2020). *RStudio: Integrated Development Environment for r*. RStudio, PBC, Boston, MA.

[35] Running, S., Mu, Q., and Zhao, M. (2017a). MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006. doi: 10.5067/MODIS/MOD16A2.006.

[36] Running, S., Mu, Q., and Zhao, M. (2017b). MYD16A2 MODIS/Aqua Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006. doi: 10.5067/MODIS/MYD16A2.006.

[37] Santaren, D., Peylin, P., Bacour, C., Ciais, P., and Longdoz, B. (2014). Ecosystem model optimization using in situ flux observations: Benefit of Monte Carlo versus variational schemes and analyses of the year-to-year model performances. *Biogeosciences*, 11(24):7137–7158.

[38] Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7):7183–7192.

[39] Thoning, K. W., Tans, P. P., and Komhyr, W. D. (1989). Atmospheric carbon dioxide at Mauna Loa Observatory: 2. Analysis of the NOAA GMCC data, 1974–1985. *Journal of Geophysical Research: Atmospheres*, 94(D6):8549–8565.

[40] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

[41] Winfree, A. T. (2001). *The Geometry of Biological Time*. Interdisciplinary Applied Mathematics. Springer-Verlag, New York, second edition.

[42] Wu, Z., Huang, N. E., Long, S. R., and Peng, C.-K. (2007). On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences*, 104(38):14889–14894.

[43] Zobitz, J. (2013). Forest Carbon Uptake and the Fundamental Theorem of Calculus. *The College Mathematics Journal*, 44(5):421–424.