Fitting the Multitemporal Curve: A Fourier Series Approach to the Missing Data Problem in Remote Sensing Analysis

Evan B. Brooks, Valerie A. Thomas, Randolph H. Wynne, *Member, IEEE*, and John W. Coulston

***Abstract*—With the advent of free Landsat data stretching back decades, there has been a surge of interest in utilizing remotely sensed data in multitemporal analysis for estimation of biophysical parameters. Such analysis is confounded by cloud cover and other image-specific problems, which result in missing data at various aperiodic times of the year. While there is a wealth of information contained in remotely sensed time series, the analysis of such time series is severely limited due to the missing data. This paper illustrates a technique which can greatly expand the possibilities of such analyses, a Fourier regression algorithm, here on time series of normalized difference vegetation indices (NDVIs) for Landsat pixels with 30-m resolution. It compares the results with those using the spatial and temporal adaptive reflectance fusion model (STAR-FM), a popular approach that depends on having MODIS pixels with resolutions of 250 m or coarser. STAR-FM uses changes in the MODIS pixels as a template for predicting changes in the Landsat pixels. Fourier regression had an R^2 of at least 90% over three quarters of all pixels, and it had the highest $R^2_{\text{Predicted}}$ values (compared to STAR-FM) on two thirds of the pixels. The typical root-mean-square error for Fourier regression fitting was about 0.05 for NDVI, ranging from 0 to 1. This indicates that Fourier regression may be used to interpolate missing data for multitemporal analysis at the Landsat scale, especially for annual or longer studies.**

***Index Terms*—Data fusion, disturbance, harmonic analysis, interpolation, phenology, time series.**

I. Introduction

THE COLLECTION of Landsat scenes dating from 1972 is one of the largest continuous freely available satellite records of the Earth's surface. At 30-m pixel resolution, Landsat imagery is used in a variety of moderate- and broad-scale applications, including change detection in land use/land cover (LU/LC) classes [1] and ecosystem monitoring [2]–[4]. The continuous record, freely available to the public since 2009, has paved the way for a new level of time series analysis that can capitalize on high spatial and temporal sampling. However, there is a nominal 16-day gap in scenes for each satellite, and many scenes are at least partially obscured by cloud cover [5]–[7]. Thus, methods are needed to facilitate multitemporal analysis of Landsat data. These methods would ideally be robust and easily implemented and would minimize sources of error in their implementation.

Manuscript received July 22, 2011; revised November 8, 2011; accepted December 11, 2011. Date of publication February 21, 2012; date of current version August 22, 2012. This work was supported by the USDA Forest Service Cooperative Agreement with Virginia Tech (Grant No. 10-CA-11330145-158).

E. B. Brooks, V. A. Thomas, and R. H. Wynne are with the Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060 USA (e-mail: evbrooks@vt.edu; thomasv@vt.edu; wynne@vt.edu).

J. W. Coulston is with the USDA Forest Service Southern Research Station, Forest Inventory and Analysis Unit, Knoxville, TN 37919 USA (e-mail: jcoulston@fs.fed.us).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2012.2183137

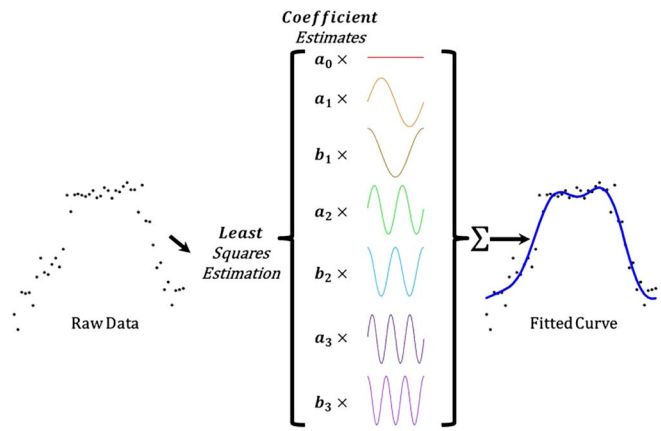

Fig. 1. Concept of Fourier regression.

Harmonic analysis using Fourier series appears to be an ideal way to facilitate multitemporal analyses using Landsat data, with demonstrated prior efficacy using coarser resolution data such as MODIS and AVHRR, particularly for phenological studies [8]–[14]. Fourier regression analysis has also been applied in areas of health [15] and land development [16].

Fourier series are superimposed sequences, over an interval of time, of a constant with sines and cosines of increasing integer multiples of the original frequency based on the time interval. The constant is called the *mean* of the series, and the pairings of sine and cosine at the specified frequencies are called the *harmonics* of the series. Fourier series can be tailored to any period length, baseline, and amplitude. As the number of harmonics used increases, the Fourier series can converge to any smooth periodic function. Fig. 1 shows the concept of using Fourier series to estimate the underlying curve in a periodic time series.

Fourier series have been shown to be useful in classification of vegetation types [12] and in the estimation of phenological markers such as start of season, peak of season, end of season, and photosynthetic activity over the growing season [8]–[11], [14]. These methods are largely focused on time series of the normalized difference vegetation index [17], or NDVI, and other similar indices, but they have not been applied to

0196-2892/$31.00 © 2012 IEEE

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

Landsat-scale data and general spectral bands. A thorough review of applications of Fourier series can be found in [14] and [10]. Some of the key results and concerns from the literature can be summed up as follows.

1) The mean and first two harmonics cover most of the variation in the data [9], [10].
2) Higher order harmonics are often needed for classifying vegetation types on a more subtle level [11], [12].
3) Fourier series of too high an order can swing wildly during times where data are missing, overfitting the remaining points (the "spurious oscillations" of [14]).
4) Fourier series coupled with polynomials (nonclassical harmonic methods) can employ higher order harmonics by reducing the "roughness" in the fit [14].
5) Multiple years may be employed in nonclassical harmonic algorithms to improve the accuracy and long-term trend detection [14].

The aforementioned findings all point to Fourier series as having great potential for use in multiyear Landsat analysis. In particular, they imply that, if the object is merely to generate "fill" images in the Landsat time series, then a basic fit using only the first two harmonics and mean may suffice. If the object is the estimation of vegetation time series or other finer scale applications, more harmonics may be required, although care must be taken to avoid overfitting the time series if it is sparse.

When fitting a functional curve to yearly data, there are several inherent advantages in using Fourier series. These include, but are not limited to, the following.

1) The fitted curve is periodic, provided that no polynomial terms are incorporated into the fit.
2) No ancillary data are required, reducing possible error sources.
3) Fourier terms are orthogonal, so there is a reduced chance of multicollinearity (defined as statistical linear association between assumed orthogonal terms in a regression model), provided that the data include dates from throughout the year.
4) Fourier terms are smooth, facilitating differential calculus approaches to time series analysis.
5) One can store the Fourier regression coefficients in raster form instead of generating images for each day of the year.

Another approach to "filling in the gaps" for Landsat coverage is the spatial and temporal adaptive reflectance fusion model, or STAR-FM [18]. Instead of a periodic approach, STAR-FM relies on the inclusion of MODIS imagery to supplement the Landsat scenes. As MODIS has a daily temporal resolution, this can provide sequences of Landsat-scaled scenes. However, MODIS spatial resolution is at best 250 m. This raises issues of accuracy in heterogeneous regions where MODIS pixels are frequently mixed [18], [19], although the enhanced ESTAR-FM [19] is designed to address this concern. Additionally, MODIS scenes are just as susceptible to cloud issues as those of Landsat. Nevertheless, STAR-FM has been shown to perform well, particularly for short (intraannual) periods of time in homogeneous areas [18], [19].

The primary aim of this paper is to demonstrate the use of a Fourier regression algorithm in comparison with STAR-FM. In particular, the objective is to check that Fourier regression may be used in lieu of STAR-FM, particularly for annual or interannual analysis of Landsat-scale scenes. One might expect Fourier series, using only 30-m Landsat pixels, to be less impeded by use in heterogeneous regions than STAR-FM, which depends on both 30-m Landsat pixels and 250-m or larger MODIS pixels. Furthermore, since no additional data are required and since the Fourier series is characterized by its coefficients, the computational and storage costs should be far less than those for using a fusion algorithm like STAR-FM. The primary objectives of this paper are the following: 1) to compare Fourier regression accuracy, in terms of fit and prediction, to STAR-FM in homogeneous regions and 2) to quantify the accuracy in heterogeneous regions. With a favorable comparison, remote sensing researchers using Landsat data will have access to another method for enabling multitemporal analysis, one that is particularly well suited to multiyear research. To demonstrate this multiyear aspect of the method, the secondary objective of this paper is to perform Fourier regression on the study area using scenes from several years. The results of this secondary analysis are compared to those of the single-year version.

II. Data

For this paper, the objective is to compare the algorithm shown in Part III with STAR-FM. To further compare the accuracies of the algorithms in different landscape types, two study areas were chosen, shown in Fig. 2. The areas are both in central North Carolina. One of them, Greensboro, is primarily urban and suburban and is sharply heterogeneous. This area is 9.4 mi by 8.8 mi (15.1 km by 14.2 km), with an area of 82.3 mi^2 (213.3 km^2). The other, the eastern area of Chatham County including east of Pittsboro, consists primarily of forested and agricultural land with a couple of lakes. This region is smaller than the Greensboro one, with dimensions of 6.9 mi by 6.6 mi (11.1 km by 10.6 km), with an area of 45.4 mi^2 (117.6 km^2). The areas of the land cover classes in this Pittsboro area are large and continuous enough to make this area fairly homogeneous. Between them, the areas include a variety of land cover classes and basic vegetation types typically found in the eastern U.S., with varying degrees of heterogeneity. This makes them suitable for testing the Fourier series, especially for forestry and classification applications.

Both study areas are from path/row 16/35 in Landsat and H/V 11/5 in MODIS. The areas are relatively proximate to control for the effects of weather and external conditions across the two areas. In all, 17 Landsat scenes were downloaded. Both MODIS Terra daily data (MOD09GQ, 346 scenes) and MODIS Terra 8-day composite images (MOD09Q1, 43 scenes) were downloaded. These scenes had a spatial resolution of 250 m. The study time for the STAR-FM comparison is the year 2001. At this time, both Landsat 5 and Landsat 7 were in operation, jointly providing images at a nominal 8-day interval. Also, at this time, the scan line corrector for Landsat 7 was still functioning. For the multiyear application, additional Landsat scenes were acquired from years 1998 to 2002 over the same

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

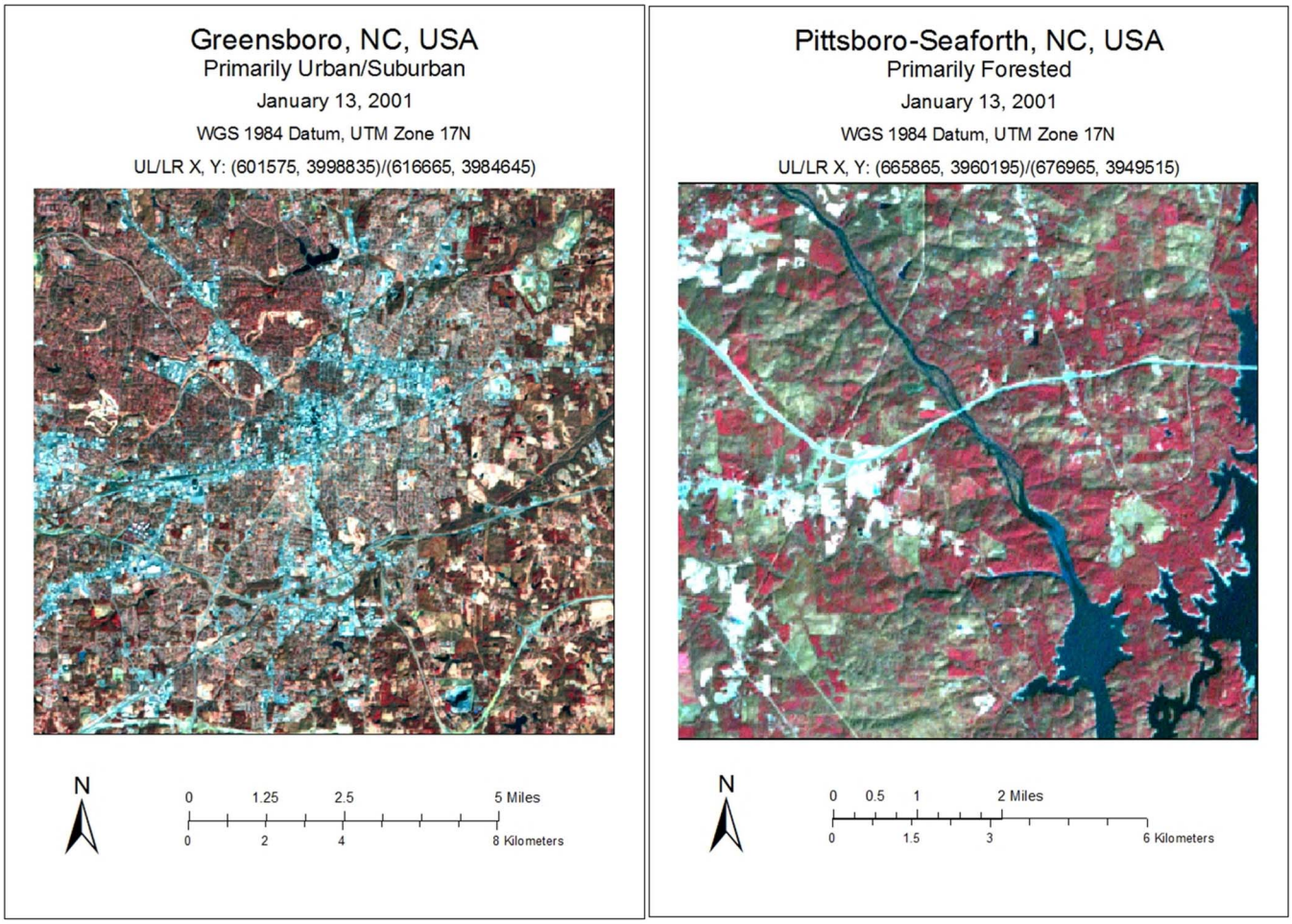

Fig. 2. Study areas. Colors for both images are Landsat bands 4/3/2 in R/G/B.

study areas. For purposes of profiling and controlling results by LU/LC classification, the National Land Cover Data (NLCD) set from 2006 was used to assign each pixel to a class. The goal of this paper is not to demonstrate an alternative classification method, although it would certainly be feasible to use Fourier regression-derived curves to aid in classification. The purpose of adding NLCD data was to allow for the possibility of controlling results by LU/LC class, in the event that one class fared particularly well or poorly. Because the MODIS scenes include bands in the red and near infrared (NIR) and are geared toward vegetation indices, it was appropriate to use the NDVI in this study. However, the algorithms are designed and intended for use in any spectral band or index. Accordingly, the Fourier regression algorithm was also run on six spectral bands from Landsat. All LU/LC classes were used in the subsequent analysis.

Preprocessing for Landsat scenes included atmospheric correction to surface reflectance using LEDAPS [20] as well as dark object subtraction in an effort to calibrate the images from the two Landsat satellites. Dark object subtraction using band minima was chosen because the year-long nature of the data rendered histogram matching ineffective due to seasonal vegetation changes. Preprocessing for MODIS included resampling, subsetting, and reprojecting the images via the MODIS Reprojection Tool [21] into Landsat scale and projection, converting the 250-m pixels to 30-m pixels.

III. Method

A. Review of STAR-FM and ESTAR-FM

Since the primary purpose of this paper is to compare the results of the Fourier regression approach with the results of STAR-FM, a review of the latter algorithm may be helpful. STAR-FM employs a sequence of linear transformations and regressions across a moving window centered on the pixel in question in order to predict the pixel's brightness value (in whatever band). Specifically, following the lead of [18] and [19], consider a situation in which we have a "fine" image at the Landsat scale of 30 m for time t_0, denoted F_0, and assume that we have "coarse" images from the MODIS scale of 250 m for times t_0 and $t_1 > t_0$, designated C_0 and C_1. The inequality may be reversed without losing generality, i.e., one can make a prediction using an input pairing after the predicted date as well. When reprojected to Landsat resolution and coordinates, we can pick out a specific pixel for each image at the corresponding location x, y from the fine image. Then, STAR-FM makes the fundamental assumption that, for any

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

given band B, the only real difference between the brightnesses of the pixels at time t_0 can be modeled by a linear bias, i.e.,

$$F_0(x, y, t_0, B) = a \times C_0(x, y, t_0, B) + b. \tag{1}$$

Using linear regression to estimate the coefficients, the algorithm then applies those estimates (designated $\hat{a}$ and $\hat{b}$) to reverse-engineer estimated values for the would-be Landsat scene at time t_1 via

$$F_{\text{predicted}}(x, y, t_1, B) = \hat{a} \times C_1(x, y, t_1, B) + \hat{b}. \tag{2}$$

In actuality, STAR-FM applies a weighted average of such estimates based on nearby pixels deemed similar to the target pixel x, y. The weights are based on three measures of the nearby pixels with respect to the target: the spectral difference between the brightness values in all the bands, the temporal distance between the pixel dates, and the spatial distance between the pixels. The estimation process can be further improved by adding a second "input pair" of fine and coarse images at another time ($t_2 > t_1 > t_0$).

It is easy to see why STAR-FM has reduced accuracy when the coarse pixels are mixed as the assumption of uniformly linear bias is called into question. ESTAR-FM [19] addresses this concern by delving into the two-input-pair model, where each coarse pixel is a weighted average of the finer pixels comprising it. ESTAR-FM then estimates pixel-specific ratios suggested by the linear changes in the two fine pixels around the prediction date. In the final model, this ratio is added into the formula for the prediction from STAR-FM, further tailoring the prediction to specific observed changes in the fine pixels.

One weakness of both STAR-FM and ESTAR-FM that emerges is that, when modeling year-round changes, particularly when large blocks of data are missing over seasonal changes, the assumption of linear change from one date to the next may not be justified. It is precisely in situations like this, where a straight line segment seems insufficient to model the change, that we would prefer to see a smoothly undulating curve filling the gaps.

B. Basic Algorithm

In the primary objective of this paper, we explore a method using only harmonic terms for a single year, although variations and other Fourier regression-based methods in the literature could ostensibly be used. The motivation for choosing one year is to facilitate comparison with STAR-FM. The goal here is to compare accuracy of prediction with STAR-FM, but it is also desirable to use enough harmonics to demonstrate the utility in LU/LC classification as well. To address the secondary objective of this paper, we use five years' worth of Landsat data in a standard Fourier regression context. This is done to compare the results with those of the single-year analysis.

An explanation of the Fourier regression algorithm follows, akin to that found in [22]. For a pixel p measured at times $\boldsymbol{t} = (t_1, t_2, \ldots, t_d)$ to have brightness values across a spectral band b, given by the time series $\boldsymbol{b} = (b_1, b_2, \ldots, b_d)$, we generate linearly interpolated fill points for gaps larger than a specified threshold g, producing combined vectors $\boldsymbol{t}^* = (\boldsymbol{t}, \boldsymbol{t_{fill}})$ and $\boldsymbol{b}^* = (\boldsymbol{b}, \boldsymbol{b_{fill}})$. The time vector may be assumed to be rescaled to the interval $[0, 2\pi]$ without loss of generality. The linear interpolation prevents the models from producing nonsensical values in fitting the relatively sparse sections of the time series. Note that the resulting dates are not necessarily evenly spaced since the object here is Fourier regression and not a transform. Further note that the linear interpolation is intended to be used for gaps considerably larger than the typical interval between Landsat data points as the goal is to suppress wild oscillations that result from large gaps in the data. If these gaps cross seasonal changes or cover major likely features of the time series, then the Fourier regression algorithm will likely have reduced accuracy, just as STAR-FM may. The major distinction to be drawn here is that the linearly interpolated points are used as supplemental training data for the regression algorithm, as opposed to being the predicted values in and of themselves.

Once the gaps are filled, we use least squares estimation to estimate harmonic coefficients; that estimate is denoted here by $\boldsymbol{a}$. In particular, we generate a model matrix using n harmonics as

$$T = (\mathbf{1} \quad \sin(\boldsymbol{t}^*) \quad \cos(\boldsymbol{t}^*) \quad \cdots \quad \sin(n\boldsymbol{t}^*) \quad \cos(n\boldsymbol{t}^*)). \tag{3}$$

Then, the coefficient estimates can be obtained using the usual least squares method of

$$\boldsymbol{a} = (T'T)^1 T'\boldsymbol{b}^* \tag{4}$$

where T' is the transpose of T. Each pixel and each spectral band for that pixel have their unique set of coefficients, so the output of the algorithm is a raster of coefficients. The more harmonics the user desires, the more layers are in the output raster. Note that there is a relationship between the choice of g and the values of n. Smaller values of g result in more fill points being generated, so small g values support a higher number of harmonics.

In order to run this paper's algorithm for Fourier regression, the user must do the following.

1) Input the Landsat image files (preferably cloud-free) that are rectified and subsetted to the same size.
2) Specify the number of harmonics desired (more harmonics = more detailed fit $\rightarrow$ greater chance of overfitting sparse data and getting "spurious oscillations").
3) Specify the gap threshold (the smaller the threshold is, the more the Fourier regression fit will resemble a linear interpolation).

The algorithm works over each discrete time series of the entire area at once, outputting multiple rasters of coefficients.

C. Specific Application

Preprocessing of the Landsat scenes was performed in ERDAS Imagine 2010 and R version 2.11.1 [23]. Imagine was used to subset and combine the images, and R was used, applying both the caTools [24] and abind [25] libraries to perform dark object subtraction and to generate binary files. Preprocessing for the MODIS images included using the MODIS Reprojection Tool [21] to subset, resample, and reproject the

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

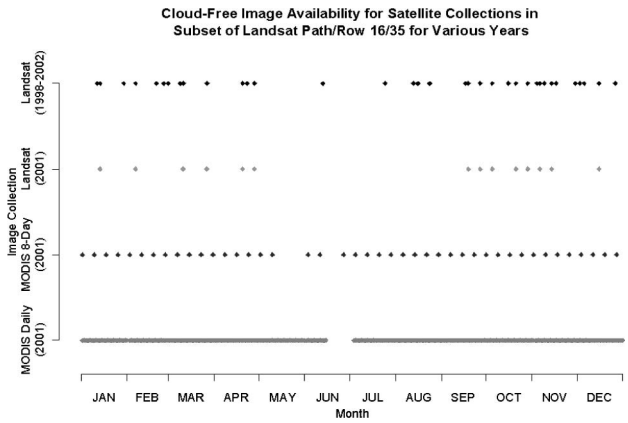

Fig. 3. Distribution of image dates for Landsat and MODIS.

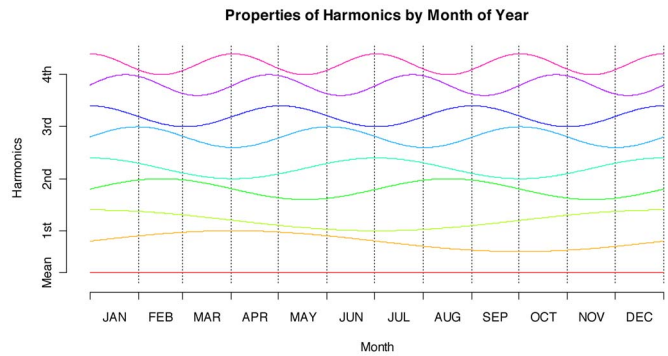

Fig. 4. Intraannual trends that may be captured by different harmonics in Fourier regression.

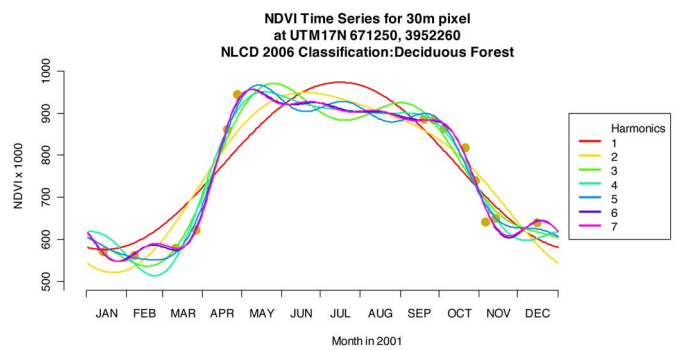

Fig. 5. Effects of increasing the number of harmonics at a sample pixel.

images using nearest neighbor resampling. R was then used to generate binary files. R was used, using custom code, to perform the Fourier regression algorithm. STAR-FM was implemented using Linux source code obtained from NASA [26]. For STAR-FM, the default values in the input files were used.

The specific dates in 2001 (as days of the year) for each of the image time series are shown in Fig. 3, as are the dates of the 1998–2002 data, shown as days of the year for compactness's sake. These dates apply to both scenes as they are subsets of the same base image. The most obvious feature of the data is that Landsat is missing values from May to mid-September. This has some interesting ramifications for the analysis since the quality of the Fourier regression depends in some sense on whether enough representative points are available. In the multiyear analysis, this problem is resolved by using multiple years and by achieving a more even spread of points over the course of the days of the year. In the interest of comparing with STAR-FM, only one year is used due to the considerable processing requirements of running STAR-FM on daily MODIS images over the course of multiple years. Because the primary object of comparison is NDVI time series, for the type of land cover considered here, the missing summer months can ideally be represented as long as data are available for the peak of the greening in late spring and the start of senescence in autumn for the classes of land cover considered in this paper.

Due, in part, to the large gap in the Landsat time series, the gap threshold for the single year analysis was chosen to be $g = 32$ days. As a result, six fill points were generated for the summer values on a line interpolating the endpoints of the gap. If the gap was smaller, a smaller value of g might have been chosen, but to do so in this case would invite excessive fitting to interpolated data. A larger value would invite nonsensical undulations in the fit. It would be possible to develop an automated rule based on the distribution of dates across the year and a desired number of harmonics to determine a value for g, but this was beyond the scope of this paper in applying Fourier regression to the particular study areas. For the multiyear analysis, g was set at 365 days, guaranteeing that there would be no linear interpolation at all in the multiyear analysis.

The number of harmonics n was chosen to reflect the nature of NDVI (vegetative index) and the object of this paper (comparison of fit and predictive robustness with STAR-FM). With 12 months in a year, the first four harmonics were chosen to allow for variation on a month scale. This concept is shown in Fig. 4, where the basic harmonics are shown as a constant mean and increasing pairs of sines and cosines with unit amplitude. The months are delineated by dashed lines, and by observing the harmonics' behavior between each pair of lines, one can see that each month has its own unique "harmonic address." There are biannual trends represented in the first harmonics, triannual trends in the second harmonics, etc. The fourth harmonics allow changes on a monthly scale, although using the fourth harmonics alone would force such changes to recur every other month. By using all of the harmonics together, a detailed curve for the year can thus be obtained. In more sparse data sets, there is a need to choose fewer harmonics to avoid generating misleading undulations as use of four harmonics will compel the curve to detail monthly changes, even if there are not enough data to support them.

As a more concrete example, Fig. 5 shows the effects of increasing harmonics on a sample time series. With only the first harmonic, the rough shape of the time series is outlined, but little else is fitted well. Increasing the harmonics shifts the peak of the curve toward the last known point before the gap in April. By the time six or seven harmonics are used, minor details in the raw time series are fitted more closely. The effect of the linear constraints in the gap can be seen in the higher harmonics as well. For comparison, the typical deciduous forest NDVI time series can be thought of as "mesa-like" [27],

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

with rapid rises and drops during the spring greening and fall senescence, respectively. The curve in Fig. 5 certainly fits that description.

IV. Analysis

A. Comparison to STAR-FM

Since the primary objective is to compare the accuracy of Fourier regression with STAR-FM, some sort of validation data are needed. In this case, only cloud-free Landsat scenes were selected, and all algorithms were run with the assumption that every scene truly reflected conditions on the ground. The Fourier regression was performed on the NDVI values derived from the Landsat scenes (augmented with the six interpolated missing values for the summer), as well as on each of the six spectral bands independently. STAR-FM was run on both the daily and 8-day MODIS images, using two input pairs where possible. STAR-FM accepts inputs on a band-by-band basis, so it was run for both the red and NIR bands (MODIS bands 1 and 2, respectively) separately before combining the outputs to compute the NDVI.

Fourier regression was run only on the Landsat data, and STAR-FM was run using both the Landsat and MODIS reprojected data. Because the objective was to produce a good fit on vegetation index data over the course of 12 months, the Fourier regression was run, using the six interpolated fill values over the summer months, with four harmonics (potentially allowing month-by-month variations in the basic curve). Fourier regression was applied to the Landsat NDVI and to each of the six Landsat bands independently. Fourier regression was not used on any of the MODIS data. STAR-FM was applied using both the daily MODIS imagery and the 8-day composites in conjunction with the Landsat scenes, effectively allowing for a three-way comparison between Fourier regression and the two shades of STAR-FM.

In all cases, only algorithm outputs corresponding to the known Landsat dates were considered for the checking of fitting and predictive accuracy. The interpolated values were used in stabilizing the curve throughout the summer months, but those values were not checked for accuracy as there were no Landsat data available to check them against. In order to check the predictive accuracy, deleted residuals were calculated. In the general regression context where $\mathbf{y}$ is the response vector, for any point i, let $\hat{y}_{(i)}$ be the predicted value for the ith point from the model generated by all points except the ith point. Then, the deleted residual is defined as $\hat{e}_{(i)} = y_i - \hat{y}_{(i)}$. Deleted residuals are desirable here because the interest is in the algorithms' abilities to predict values for missing dates, in addition to accuracy of overall fit. For the Fourier regression, this was achieved by removing each point one at a time and then by implementing the fill interpolation each time before estimating coefficients and calculating the deleted residual. For STAR-FM, input pairs were used around the target date, using only the MODIS scene at the missing date to make the deleted prediction.

For accuracy of fit, the standard measures of root-mean-square error (RMSE) and R^2 may be used. For prediction, summing the squares of the deleted residuals gives the prediction sum of squares (PRESS) statistic $\sum_{t=1}^{d} \hat{e}_{(i)}^2$. These PRESS statistics can then be compared (lower values imply greater overall accuracy), or alternately, their corresponding predicted R^2, denoted by $R^2_{\text{Predicted}}$, can be compared instead via the formula

$$R^2_{\text{predicted}} = 1 - \frac{\text{PRESS}}{\text{Sum of Squares Total}}. \tag{5}$$

Due to the large number of pixels analyzed, violin plots [28] are a useful way to describe the results without resorting to single-number summary statistics. A violin plot may be thought of as the hybrid child of a boxplot and a continuous histogram. While the median, quartiles, and trimmed extremes are preserved as in a boxplot, a density estimation method is applied to the data to generate a continuous curve. This curve is rotated and given symmetry, producing a "double" effect. The thickest parts of the plot correspond to the parts of the distribution which are most densely populated. These plots were used in summarizing the resulting statistics for the scenes.

B. Multiyear Analysis

In the case of the second objective, for simplicity, only the Landsat NDVI time series were considered. As in the single-year analysis, only cloud-free dates were chosen. The extra years did fill in the summer values missing from the 2001 data, as was shown in Fig. 3. In order to perform Fourier regression, the dates for the scenes were converted into days of the year for the combined meta-year made from superimposing all the days of the year from the data. Additionally, the dates were also recorded as days counted from the beginning of 1998, allowing for the possibility of including polynomial terms in addition to the Fourier regression terms in the event a researcher wishes to use nonclassical harmonic analysis. Such a method would be well suited to looking for interannual growth or decline trends over time.

As mentioned earlier, the multiyear analysis was performed using no linear interpolation whatsoever as the gaps between points in the meta-year were much smaller. This analysis is pure harmonic regression. As before, R^2 values were calculated from the resulting fits. These values were compared to the R^2 values obtained from applying Fourier regression to the single-year data.

V. Results

A point worth noting is that the results which follow represent all of the LU/LC classes grouped together. Stratification along LU/LC lines was performed, but generally, the results were so similar that the additional separation was not represented here. This is true for both the fitted and predictive results, as well as for the interannual analysis. The major exception to this rule was water pixels, which had wildly diverging NDVI patterns depending on the dates.

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

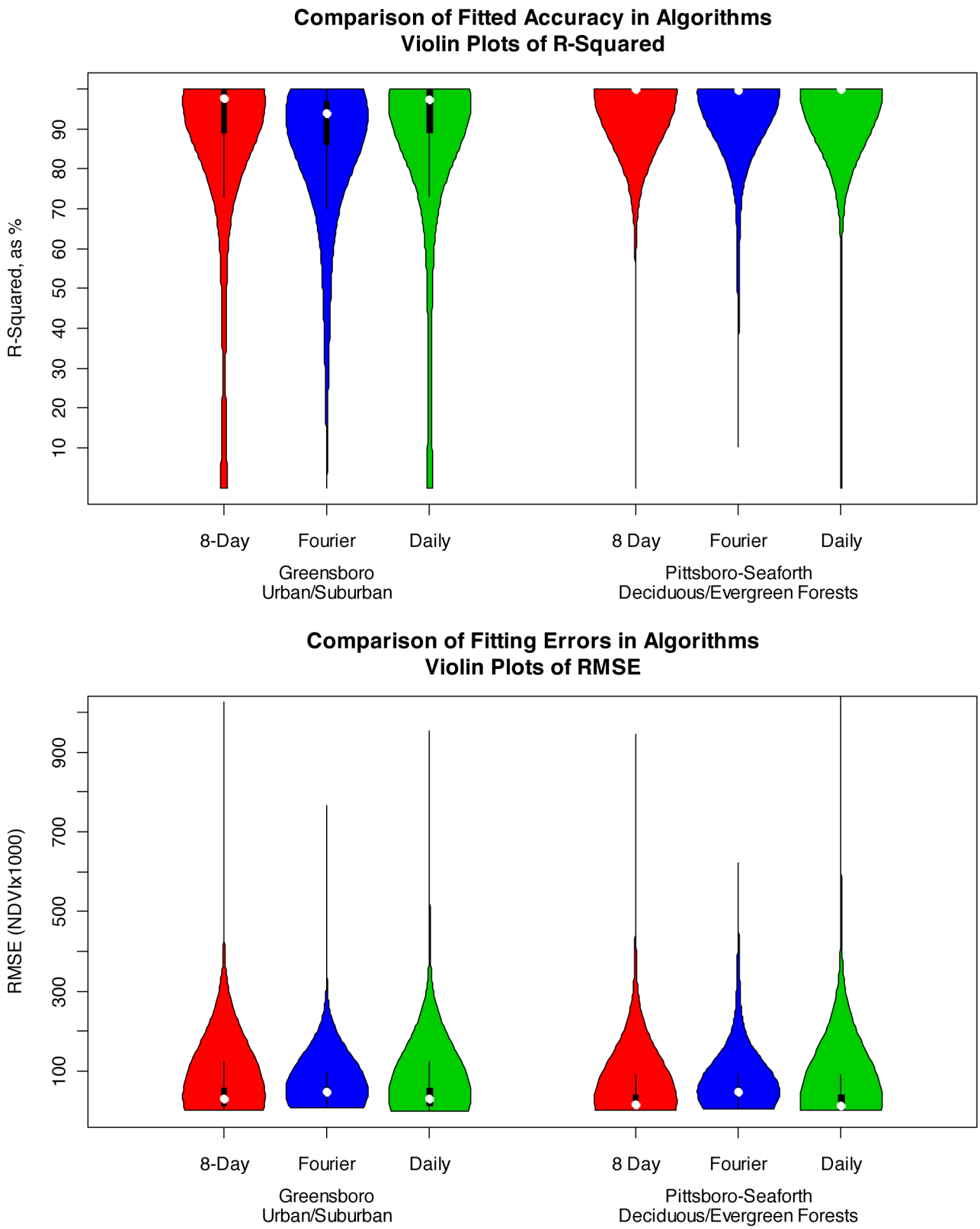

Fig. 6. Distributions for fitting statistics.

A. Fitting

The main results of fitting accuracy, the fitted R^2 and the RMSE, are shown in Fig. 6. The units for the RMSE are in NDVI $\times$ 1000, so an RMSE of 50, for example, implies that the standard deviation of observed values about the fitted values is about 0.05. It must be stressed that, in the calculation of the resulting statistics, only points corresponding to the known

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

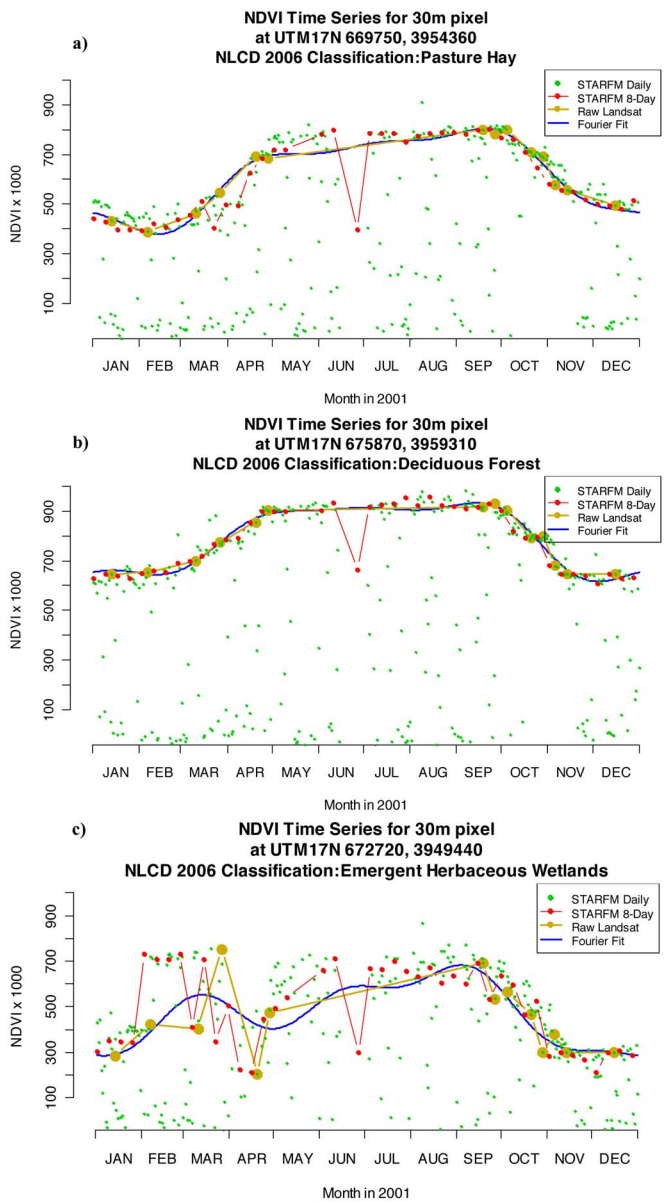

Fig. 7. Example time series and algorithm fits.

Landsat dates were considered. The interpolated fill points generated by the algorithm served to stabilize the curve but were not used directly to calculate R^2 or RMSE.

Fig. 7(a) shows a sample pixel's time series and the results of all three algorithms fitted to it. The cloud of STAR-FM points from the daily MODIS images illustrates the issue of cloud cover on MODIS pixels. A good number of pixels (about 46% of them) fall well outside the trend depicted by the raw data and the other algorithms. These dates are unsuitable for use in any sort of interpolation. Even the 8-day composite images are susceptible to this issue, as shown by the outlier point in late June. This is not a failing of STAR-FM, and these points were not used in calculating the fit and predictive statistics due to the fact that the Landsat scenes were chosen to be cloud-free, but it does illustrate the issue of cloud cover and how fusion methods must contend with it. On the other hand, the STAR-FM fits rise beyond the Fourier regression fits over the summer months, indicating that the last known Landsat point was before the greening for that year had been completed. The Fourier regression curve faithfully follows the linearly interpolated Landsat data, owing largely to the constraints placed on it by the choice of the gap threshold. In terms of comparison to the Landsat points, Fourier regression has the highest R^2 value, but it could have benefitted from at least one date in summer.

Fig. 7(b) shows a deciduous forest pixel, one in which the NDVI appears to remain fairly linear over the course of summer. The time series details a vegetative curve over the course of a year without disturbance. The trend appears smooth and undulating with a yearly period, precisely the sort of situation for which Fourier regression is suited. As in the previous figure, Fourier regression fits the known Landsat data tightly without any outlier issues, and it does so despite the fact that four months' worth of data are missing. The accuracy of the summer months—in which the fit is based primarily on the linear interpolation—cannot be determined, but the agreement of the curve with STAR-FM's output for the 8-day composites, particularly the early summer, is heartening.

Fig. 7(c) shows another pixel in which the raw Landsat series contains some problematic points. While both images were restricted to dates that were as cloud-free as possible, there were some pixels that suffered from haze or shading. This one, near a water body, also mixed land reflectance with that of the water's surface, which had a deleterious effect on the flow of the series (i.e., one of the assumptions in the Fourier regression algorithm was violated). As a result, all three algorithms suffer in accuracy, but the STAR-FM points are better able to cope with the rapid swings of the time series.

From the violin plots of Fig. 6, all three algorithms fit the Landsat data quite nicely most of the time. Note that, in the Greensboro area, the Fourier regression algorithm, while having a slightly lower median R^2 at 93.9%, also has thinner tails than the STAR-FM values and thus has a higher mean R^2 at 87.6%. The median values in the Pittsboro area are all above 99.8%, but note that Fourier regression again has the smallest "poor fit" tail of the three algorithms. The conclusion to be drawn from Fig. 6 is that all three algorithms actually fit the known data quite well.

B. Prediction

The results of the predictive comparison are shown in Fig. 8. Again, only the known Landsat dates were used in calculating both $R^2_{\text{Predicted}}$ and the predictive RMSE. Since there are hundreds of thousands of pixels in each area, any statistical comparison of means and medians will produce uninformatively significant results, but a visual inspection of the plots indicates the important features of the comparison. Clearly, all three algorithms can be perturbed by missing data, but Fourier regression seems the least perturbed of the three. In particular, in the Greensboro area, the STAR-FM algorithms suffer from higher predictive RMSEs than Fourier regression, including some truly extreme values. Generally, the algorithms did much better in the Pittsboro area, although upon checking this was not due to a difference in land cover class distribution according to the land cover assignments made by the NLCD 2006 data set.

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

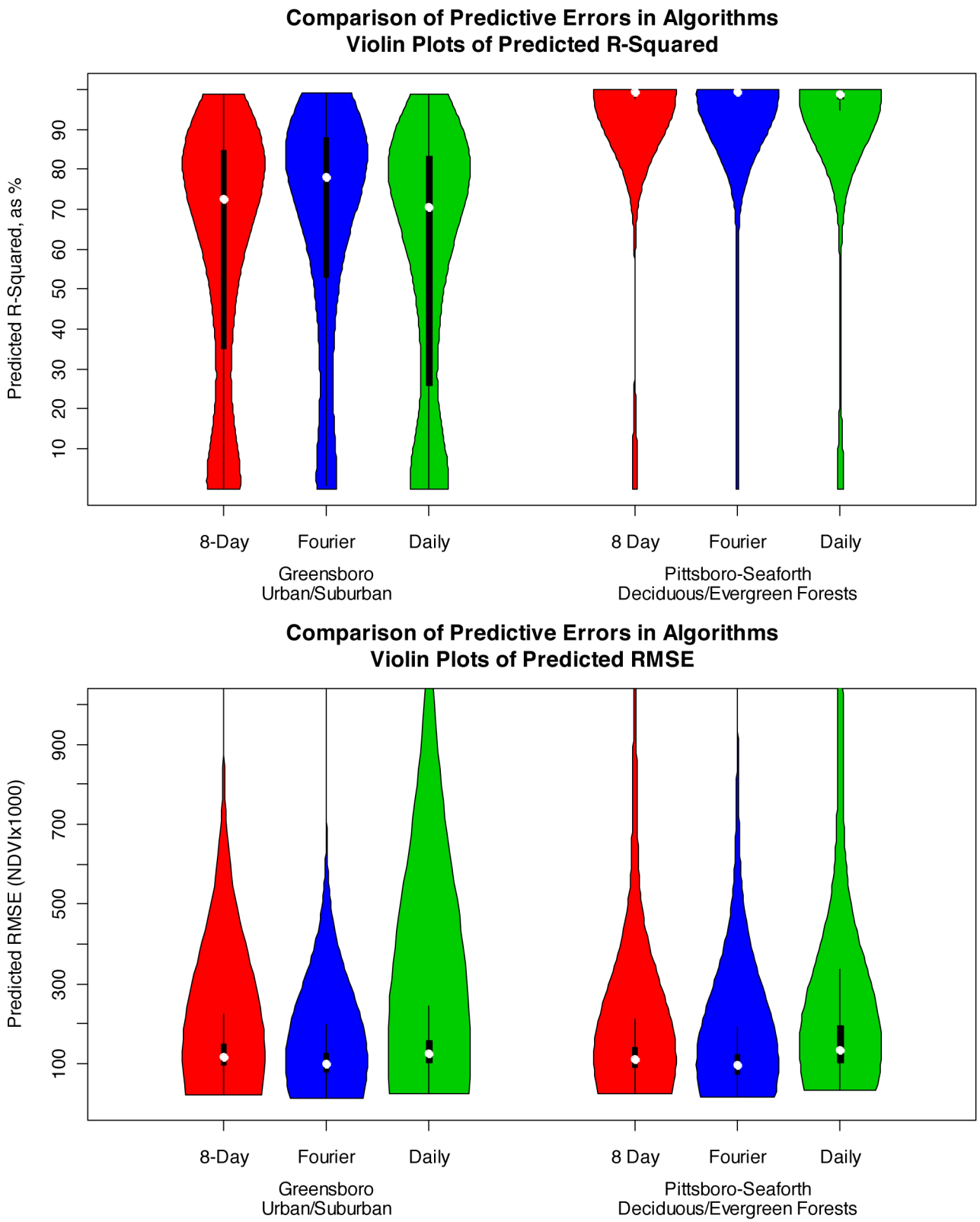

Fig. 8. Distributions for predictive statistics.

The chief conclusion to be drawn from Fig. 8 is that Fourier regression is more robust to missing data than STAR-FM, particularly in the relatively heterogeneous Greensboro area. This is somewhat surprising since STAR-FM had the benefit of ancillary data to compensate for missing values, but it speaks well for the assumption that the NDVI follows a curve that a Fourier series can appropriately model.

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

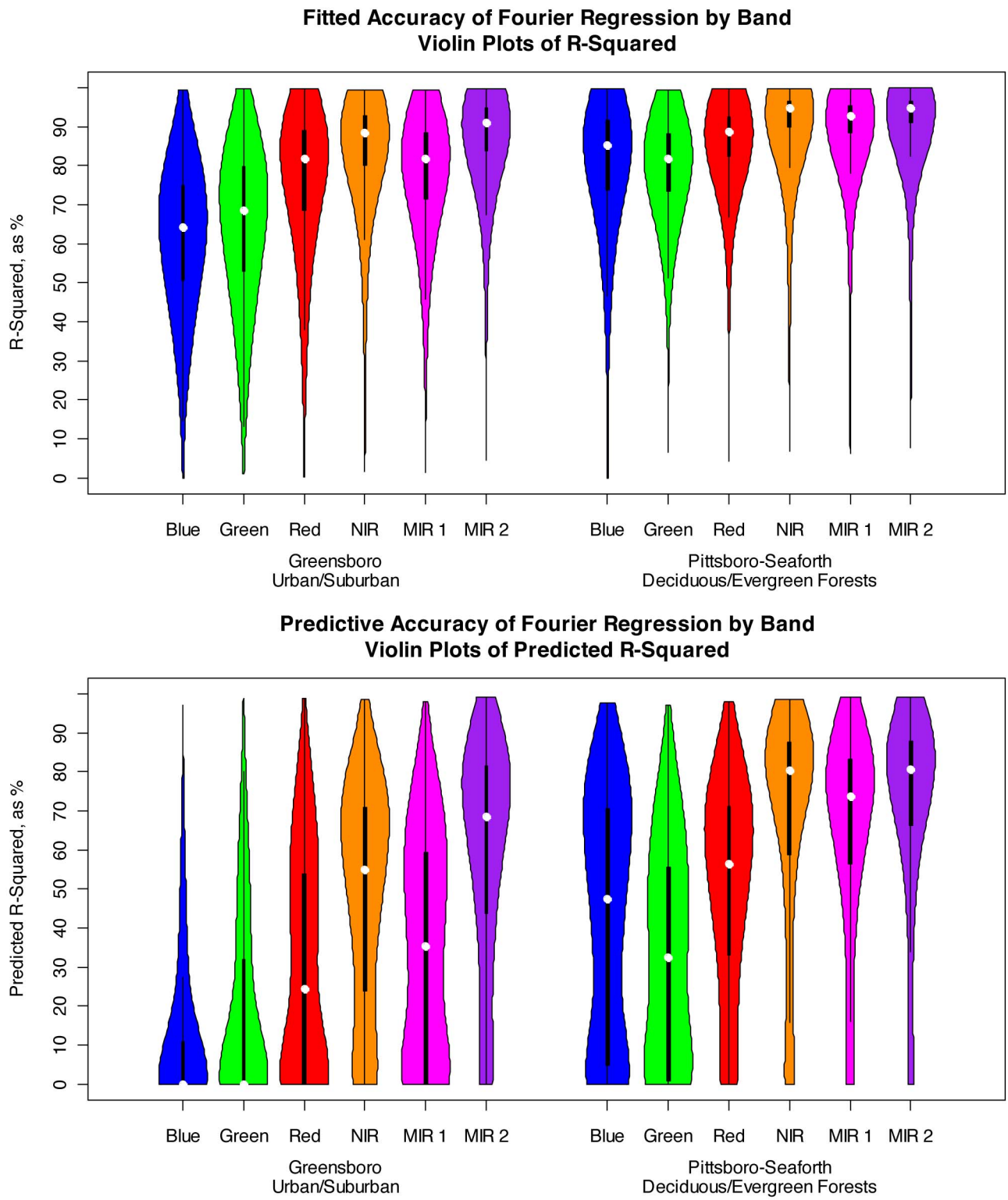

Fig. 9. Distributions of fit and predictive statistics across Landsat bands.

C. Basic Landsat Bands

In addition to comparing the Fourier regression method to STAR-FM, Fourier regression was run on six Landsat bands: blue, green, red, NIR, and two mid-infrared bands. The objective of this was to check the fitting accuracy and predictive robustness of Fourier regression when dealing with a nonindex data set, primarily for purposes of image generation or missing value imputation. The results of the analysis are shown in Fig. 9.

It is immediately clear from Fig. 9 that Fourier regression is more accurate in the infrared bands, both in terms of fit and in terms of prediction. In particular, the blue and green bands

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

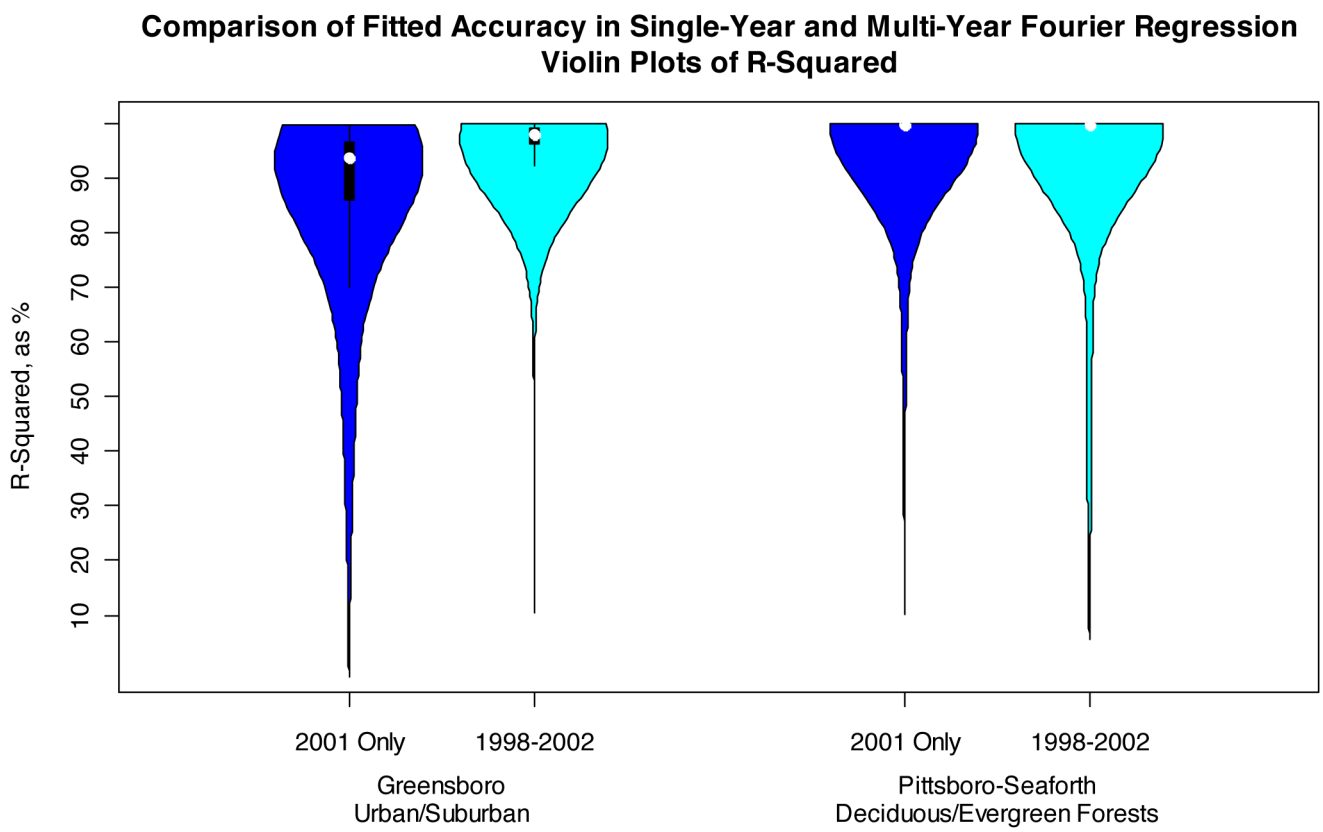

Fig. 10. Distributions for fitted comparing single-year and multiyear approaches.

suffer dramatically in R^2 to the point that the median values are below 70%. This is, in part, an artifact of the inherent band variability, as the variation in the visible bands tends to be much lower than that of the infrared bands, resulting in smaller total variation and thus lower R^2 values.

Despite the difficulties with the visible bands, it does appear that Fourier regression produces reasonably accurate facsimiles of the time series of basic Landsat bands, although those predictions are somewhat less robust than originally anticipated.

D. Multiyear Analysis and Comparison

The main results of the comparison between the 2001-only data and the data from 1998–2002 are shown in Fig. 10. The chief interpretation is that using extra years greatly improved the fitting accuracy of the method for pixels in the urban/suburban Greensboro area, while the extra years actually reduced the lower end of the fitting accuracy in the forested Pittsboro–Seaforth area. This reduction is not severe as the 20th percentile of the multiyear data is still at 99.2%, compared to the 20th percentile value of 99.7% for the 2001-only data. The vast majority of the data are well above an R^2 of 99%.

To further demonstrate the potential of multiyear analysis, information regarding a nicely fitted pixel is shown in Fig. 11. The pixel in question comes from the forested Pittsboro–Seaforth area. It is classified by the 2006 NLCD data set as deciduous forest. Fig. 11(a) shows the meta-year generated by superimposing days of the year from 1998 to 2002. It is easy to see that gaps from the 2001-only analysis are filled by values in the other years, eliminating the need to use a linearly interpolated fill algorithm before performing Fourier regression. The regression curve balances out the years' data well, resulting in a fitted R^2 of 99.8% on the known data points. In Fig. 11(b), the curve is redrawn over the course of the years, showing the way in which the yearly periodicity of the NDVI is captured by the curve. In particular, when comparing Fig. 11(b) and (c), there is a slight increasing trend in the NDVI series off the curve suggested over the years 2000–2002. This could be tested by a simple linear regression model to determine whether the coefficient is statistically significant. The residuals in Fig. 11(c) are the key to applying Fourier regression in hopes of detecting disturbances and trends over time using multiyear Landsat data.

With dates available throughout the year, there was an opportunity to see whether particular parts of the year, such as summer or winter, were better predicted by Fourier regression than other parts of the year. This was checked by calculating the residual values left over from subtracting the Fourier regression's fitted values from the known Landsat values and then summarizing the residuals across the forested Pittsboro–Seaforth study area by day of the year. It was assumed that the urban/suburban Greensboro area would have been of less interest from a phenological point of view. Fig. 12 shows the resulting interquartile ranges by day of the year.

It is very interesting to note that the one season in which both the accuracy and range around the median were minimized was summer. This could be due to a number of factors. First, the presence of snow in the winter months could have effects similar to cloud cover in calculating NDVI values, causing

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

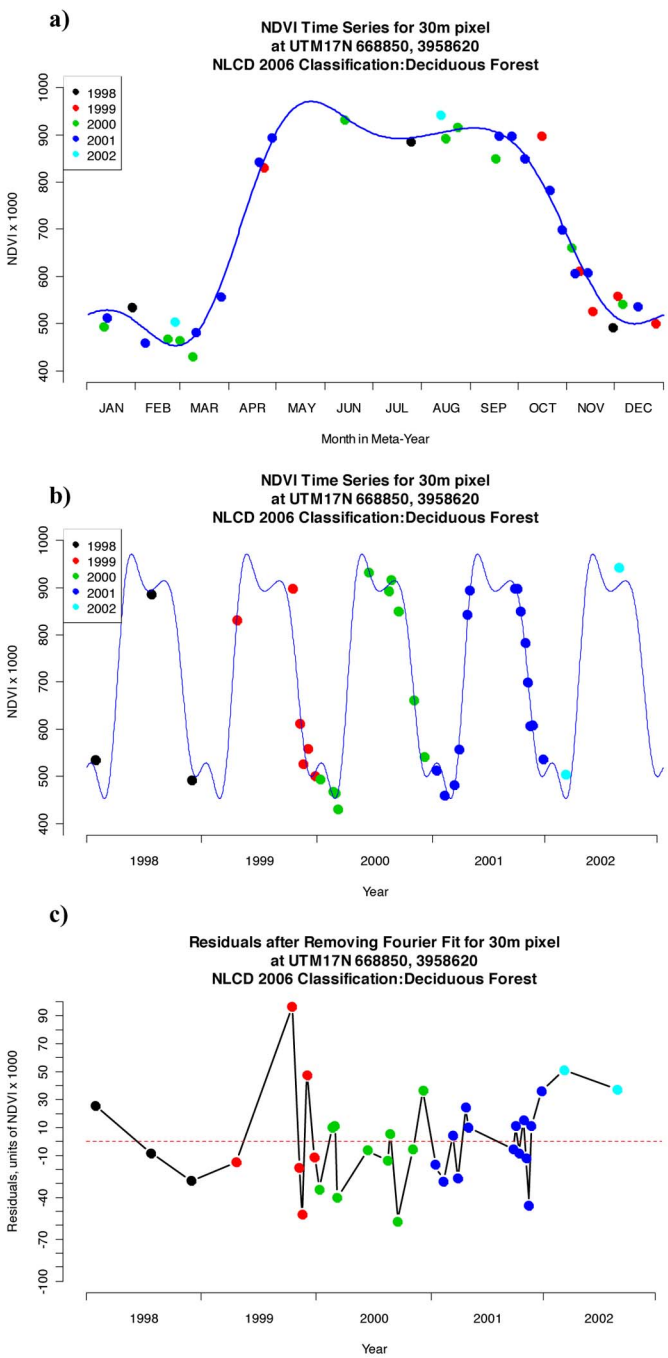

Fig. 11. Multiyear analysis of a specific pixel, with residual time series.

more error and wider variability around the expected values. Although the data had been screened to remove cloudy dates, they had not been screened to exclude all dates with snow on the ground. Second, the NDVI time series tend to peak in early summer and to remain high until fall. Any curves exhibiting this basic trend, especially where the NDVI values are near to 1 anyway, would be likely to be close to the real values by virtue of the curves' construction. That being noted, it is informative to note that the curves had higher error in the transition seasons of spring and fall than in summer. It is also worth noting that this higher error is usually not in itself great, on the order of 0.05 to 0.10 in NDVI.

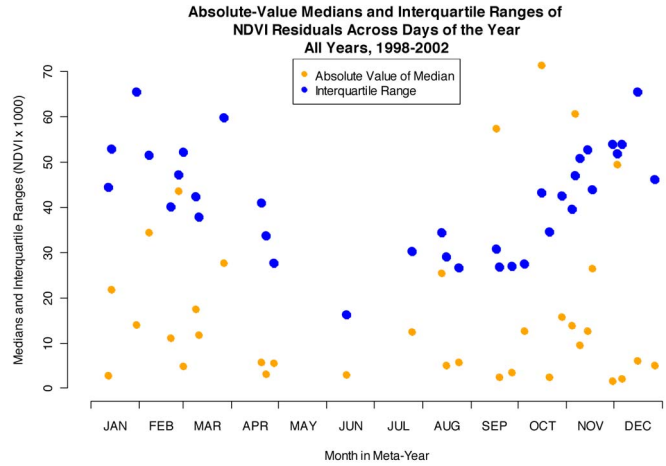

Fig. 12. Summary statistics of fitted residuals, by day of year, for the forested Pittsboro–Seaforth area.

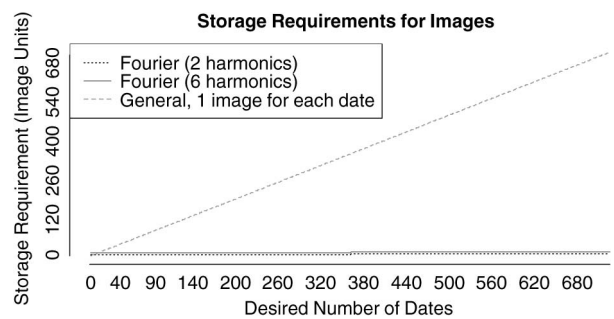

Fig. 13. Storage requirements for interpolating Landsat data throughout a desired number of dates.

VI. Conclusion

The results pertaining to the primary objective of this paper show that, for the types of land cover studied here, Fourier regression and STAR-FM are indeed comparable in the "middle ground" of a single-year analysis. In some sense, Fourier regression was put at a disadvantage in this analysis by using only a single year to train the data. In the results pertaining to the second objective, Fourier regression was performed using five years' worth of Landsat data, having the effect of improving Fourier regression's accuracy overall. The residual values left over after the interannual analysis are of considerable interest in their own right, opening avenues for change detection methods and trend observations over time for each pixel. There is potential for an "on-the-fly" disturbance detection, using previous years' data to check whether an incoming scene matches the expectation via a statistical method.

A key advantage of Fourier regression that emerged in the study was its ability to reduce storage and processing requirements. Suppose, for example, that one wished to generate data to cover a full 365-day year at daily temporal resolution for a scene comprising 1 GB of data. Instead of generating 365 individual images (and requiring 365 GB to save them), one can instead save the Fourier harmonic coefficients and use them to generate the data as needed (e.g., pixel by pixel). Even with six harmonics plus a constant, this would result in only storing 13 rasters instead of 365. If the coefficients are converted to

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON BETWEEN FOURIER REGRESSION AND STAR-FM

	Fourier Regression	STAR-FM
Advantages	• Robust, accurate prediction and fit • Reduced storage space (can save ~9 harmonic coefficients instead of ~350 prediction images) • No ancillary data • Suited for interannual studies • More harmonics = finer fit	• Robust, accurate prediction on cloud-free days • Availability of composite imagery • Able to handle sudden changes on a daily basis • Suited for intra-annual studies, especially for short duration
Disadvantages	• Must have input data at key points of curve • Harmonics limited by quantity of data • Requires at least one year of data • Produces undesirable "wiggles" • Fits poorly when pixel undergoes disturbance	• Nontrivial processing/computing requirements • Must generate images for each prediction date • Susceptible to cloud cover issues • Reduced accuracy in heterogeneous areas • MODIS has no blue band, only a blue-green

an integer format through multiplying and truncating, then the entire interpolation can be saved in only 13 GB of space. If a multiyear study was desired, one could save a couple of polynomial coefficients to account for interannual trends in land cover at the cost of only 1 or 2 GB, as opposed to another full 365 GB, further compounding the savings. Fig. 13 shows this idea. In the multiyear analysis done in this paper, no polynomial terms were added, so the storage costs for the five-year model were equal to the costs for the single-year model. From this example and the results of that analysis, one can intuit that adding more dates into Fourier regression can only improve the quality of the estimated coefficients for the model at the same storage cost.

STAR-FM is not as well suited to interannual analysis, owing to the need to generate images for each desired point in each year using multiple images to make each predicted image. Although the input images may be reused for different prediction dates, STAR-FM still requires nontrivial processing and space. Considering that Fourier regression has been shown to be comparably accurate, it does not seem efficient to use STAR-FM for this sort of analysis. On the other hand, if the time of interest is well within a single year, STAR-FM is not bound by the constraints of periodicity and does not require a full year of Landsat images to run. In such cases, STAR-FM is clearly a better choice than Fourier regression. Table I details some of the advantages and disadvantages of both methods.

The ability to use smooth curves to represent yearly Landsat data makes many different forms of analysis possible. As examples of possible applications, one may use various curve features such as integral area, maximum/minimum, and Fourier regression coefficient values as explanatory variables in regression models. These variables may then be tied to ground observations of biophysical parameters such as biomass, and from the resulting model, one may estimate biomass for a given scene from the Landsat data. If a regression model requires the minimum or maximum value of a pixel over the year, the Fourier regression curves may be used to gain estimates.

The possibilities for application of a smooth periodic curve to represent changes in brightness values over time are legion. Only a few have been touched in this paper as the goal was to demonstrate a method for making such a curve and in comparing it to STAR-FM. However, any context in which at least one year's worth of Landsat data is available may make use of Fourier regression to fill in the missing values. The ultimate conclusion of this paper is that, for the purposes of annual and interannual time series analysis of Landsat scenes, particularly in regions similar to the eastern U.S., Fourier regression is a good choice for fitting the multitemporal curve.

ACKNOWLEDGMENT

The authors would like to thank C. Blinn of the Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, for her help in implementing STAR-FM.

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] C. E. Woodcock and M. Ozdogan, "Trends in land cover mapping and monitoring," in *Land Change Science*, Gutman, Ed. New York: Springer-Verlag, 2004, pp. 367–377.

[2] S. P. Healey, W. B. Cohen, Z. Q. Yang, and O. N. Krankina, "Comparison of tasseled cap-based Landsat data structures for use in forest disturbance detection," *Remote Sens. Environ.*, vol. 97, no. 3, pp. 301–310, Aug. 2005.

[3] J. G. Masek and G. J. Collatz, "Estimating forest carbon fluxes in a disturbed southeastern landscape: Integration of remote sensing, forest inventory, and biogeochemical modeling," *J. Geophys. Res.—Biogeosci.*, vol. 111, p. G01 006, Feb. 2006, DOI:10.1029/2005JG000062.

[4] J. G. Masek, C. Q. Huang, R. Wolfe, W. Cohen, F. Hall, and J. Kutler, "North American forest disturbance mapped from a decadal Landsat record," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2914–2926, Jun. 2008.

[5] G. P. Asner, "Cloud cover in Landsat observations of the Brazilian Amazon," *Int. J. Remote Sens.*, vol. 22, no. 18, pp. 3855–3862, 2001.

[6] P. V. Jorgensen, "Determination of cloud coverage over Denmark using Landsat MSS/TM and NOAA-AVHRR," *Int. J. Remote Sens.*, vol. 21, no. 17, pp. 3363–3368, 2000.

[7] J. C. Ju and D. P. Roy, "The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 1196–1211, Mar. 2008.

[8] H. Carrao, P. Gonalves, and M. Caetano, "A nonlinear harmonic model for fitting satellite image time series: Analysis and prediction of land cover dynamics," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 4, pp. 1919–1930, Apr. 2010.

[9] G. J. Roerink, M. Menenti, and W. Verhoef, "Reconstructing cloud-free NDVI composites using Fourier analysis of time series," *Int. J. Remote Sens.*, vol. 21, no. 9, pp. 1911–1917, 2000.

[10] A. Moody and D. M. Johnson, "Land-surface phenologies from AVHRR using the discrete Fourier transform," *Remote Sens. Environ.*, vol. 75, no. 3, pp. 305–323, Mar. 2001.

[11] M. E. Jakubauskas, D. R. Legates, and J. H. Kastens, "Harmonic analysis of time-series AVHRR NDVI data," *Photogramm. Eng. Remote Sens.*, vol. 67, no. 4, pp. 461–470, Apr. 2001.

[12] R. Geerken, B. Zaitchik, and J. P. Evans, "Classifying rangeland vegetation type and coverage from NDVI time series using Fourier filtered cycle similarity," *Int. J. Remote Sens.*, vol. 26, no. 24, pp. 5535–5554, Dec. 2005.

[13] J. F. Hermance, R. W. Jacob, B. A. Bradley, and J. F. Mustard, "Extracting phenological signals from multiyear AVHRR NDVI time series: Framework for applying high-order annual splines with roughness damping," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3264–3276, Oct. 2007.

[14] J. F. Hermance, "Stabilizing high-order, non-classical harmonic analysis of NDVI data for average annual models by damping model roughness," *Int. J. Remote Sens.*, vol. 28, no. 12, pp. 2801–2819, Jun. 2007.

[15] S. I. Hay, R. W. Snow, and D. J. Rogers, "Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data," *Trans. Roy. Soc. Trop. Med. Hygiene*, vol. 92, no. 1, pp. 12–20, Jan./Feb. 1998.

[16] W. W. Immerzeel, R. A. Quiroz, and S. M. de Jong, "Understanding precipitation patterns and land use interaction in Tibet using harmonic analysis of SPOT VGT-S10 NDVI time series," *Int. J. Remote Sens.*, vol. 26, no. 11, pp. 2281–2296, Jun. 2005.

[17] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sens. Environ.*, vol. 8, no. 2, pp. 127–150, May 1979.

[18] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the Landsat and MODIS surface reflectance predicting daily Landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006.

[19] X. Zhu, J. Chen, F. Gao, X. Chen, and J. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2610–2623, Nov. 2010.

[20] J. G. Masek, E. F. Vermote, N. E. Saleous, R. Wolfe, F. G. Hall, K. F. Huemmrich, F. Gao, J. Kutler, and T.-K. Lim, "A Landsat surface reflectance dataset for North America, 1990–2000," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 68–72, Jan. 2006.

[21] J. Dwyer, J. Weiss, G. Schmidt, T. Logar, R. Burrel, G. Stubbendieck, J. Rishea, B. Misterek, S. Jia, and K. Heuser, The MODIS Reprojection Tool, American Geophysical Union, Spring Meeting, Washington, DC, 2001, abstract #U21A-24.

[22] P. Bloomfield, *Fourier Analysis of Time Series: An Introduction*, 2nd ed. New York: Wiley-Interscience, 2004.

[23] R Develop. Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Found. Stat. Comput., 2010. [Online]. Available: http://www.R-project.org

[24] J. Tuszynski, "caTools: Tools: Moving window statistics, GIF, Base64, ROC AUC, etc." R package version 1.11, 2010. [Online]. Available: http://CRAN.R-project.org/package=caTools

[25] T. Plate and R. Heiberger, "abind: Combine multi-dimensional arrays," R package version 1.3-0, 2011. [Online]. Available: http://CRAN.R-project.org/package=abind

[26] LEDAPS Tools website. [Online]. Available: http://ledaps.nascom.nasa.gov/tools/tools.html

[27] W. W. Hargrove, J. P. Spruce, G. E. Gasser, and F. M. Hoffman, "Toward a national early warning system for forest disturbances using remotely sensed canopy phenology," *Photogramm. Eng. Remote Sens.*, vol. 75, no. 10, pp. 1150–1156, Oct. 2009.

[28] J. L. Hintze and R. D. Nelson, "Violin plots: A box plot-density trace synergism," *Amer. Stat.*, vol. 52, no. 2, pp. 181–184, 1998.

Evan B. Brooks received the M.A. degree in mathematics from the University of North Texas, Denton, in 2007 and the M.S. degree in statistics from Virginia Polytechnic Institute and State University, Blacksburg, in 2010, where he is currently working toward the Ph.D. degree in the Department of Forest Resources and Environmental Conservation.

His particular research interests focus on monitoring and communication of climate change issues.

Mr. Brooks is a member of American Society for Photogrammetry and Remote Sensing, with career plans of both research and teaching in environmental and remote sensing applications of statistics.

Valerie A. Thomas received the B.Sc.(Eng) degree in environmental engineering from the University of Guelph, Guelph, ON, Canada, in 1996, the Advanced Diploma in remote sensing from the College of Geographic Sciences, Lawrencetown, NS, Canada, in 1998, and the M.Sc. and Ph.D. degrees from Queen's University, Kingston, ON, in 2001 and 2006, respectively.

In 2006, she was a Postdoctoral Fellow for Fluxnet-Canada in the Department of Geography, Queen's University. In 2007, she began working as an Assistant Professor with the Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg. Her research interests include the use of remote sensing to examine forest canopy physiology, structure, and function across space and time.

Randolph H. Wynne (M'99) received the Ph.D. degree from the University of Wisconsin, Madison, in 1995.

He is currently a Professor with the Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, where he also serves as the Associate Director of the Conservation Management Institute and Codirector of the Center for Environmental Applications of Remote Sensing. He is a member of the Landsat Science Team. His current remote sensing research focuses on algorithm development, image time series analysis of forest disturbance and recovery, ecosystem service assessment and modeling, and terrestrial ecology.

John W. Coulston received the B.S. degree in forestry, the M.S. degree in forest biometrics, and the Ph.D. degree in forest ecology and remote sensing from North Carolina State University, Raleigh, in 1996, 1999, and 2004, respectively.

He is currently the Team Leader for Methods and Techniques Research for the U.S. Forest Service, Southern Research Station, Forest Inventory and Analysis Program. His research focuses on developing and testing quantitative techniques for broad-scale inventory and monitoring of forest resources.

Authorized licensed use limited to: University of North Texas. Downloaded on December 18,2020 at 12:48:36 UTC from IEEE Xplore. Restrictions apply.