Noise reduction of NDVI time series: An empirical comparison of selected techniques

Jennifer N. Hird*, Gregory J. McDermid
Foothills Facility for Remote Sensing and GIScience, Department of Geography, University of Calgary, Calgary, Alberta, Canada T2N 1N4

ARTICLE INFO

Article history:
Received 15 July 2008
Received in revised form 18 September 2008
Accepted 18 September 2008
Keywords:
NDVI time series
Noise reduction
Vegetation phenology
MODIS

ABSTRACT

Satellite-derived NDVI time series are fundamental to the remote sensing of vegetation phenology, but their application is hindered by prevalent noise resulting chiefly from varying atmospheric conditions and sun-sensor-surface viewing geometries. A model-based empirical comparison of six selected NDVI time series noise-reduction techniques revealed the general superiority of the double logistic and asymmetric Gaussian function-fitting methods over four alternative filtering techniques. However, further analysis demonstrated the strong influence of noise level, strength, and bias, and the extraction of phenological variables on technique performance. Users are strongly cautioned to consider both their ultimate objectives and the nature of the noise present in an NDVI data set when selecting an approach to noise reduction, particularly when deriving phenological variables.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Efforts to characterize vegetation phenology (Badeck et al., 2004) and extract numerical observations related to vegetation dynamics (DeFries et al., 1995; Reed et al., 1994; Reed et al., 2003) using per-pixel time series of the normalized difference vegetation index (NDVI) are hindered by noise arising from varying atmospheric conditions and sun-sensor-surface viewing geometries (Carreiras et al., 2003; Hoblen & Fraser, 1984; Huete et al., 1999; Kobayashi & Dye, 2005; Li & Strahler, 1992). Cloud, ozone, dust, and other aerosols generally decrease near-infrared reflectance, which leads to spurious drops in the data (Coward et al., 1991; Hoblen, 1986). Off-nadir viewing and low sun zenith angles can also cause a similar effect (Gutman, 1991). Indeed, the negative bias caused by unfavourable atmospheric conditions and anisotropic bidirectional effects are a prevalent and well-recognized feature of noisy NDVI data sets. The literature contains reference to a broad variety of strategies designed to reduce the impacts of these issues (e.g., Beck et al., 2006, 2007; Bradley et al., 2007; Holben, 1986; Jönsson & Eklundh, 2002, 2004; Ma & Veroustraete, 2006; Moody & Johnson, 2001; Roerink et al., 2000; Sellers et al., 1994; Swets et al., 1999), but quantitative comparisons of alternative techniques are relatively rare. With the growing importance of NDVI time series analyses in support of research on climate change (Badeck et al., 2004), biodiversity (Hurlbert and Haskell 2003), and wildlife ecology (Hebblewhite et al., 2008), there is a strong need for a more comprehensive understanding regarding noise reduction and these data sets.


Discussions concerning the effect of individual factors (e.g., phenology, annual variation) on the application of the candidate noise-reduction techniques are infrequent in the literature. Some comparison across varying vegetation types and ecological regions was undertaken by Chen et al. (2004), where particular pixels were used to illustrate the poor performance of the Fourier-based methods over particularly asymmetric phenological profiles (i.e. NDVI time series). Ma and Veroustraete (2006) applied three separate methods to a number of test pixels covering various vegetation community types, but did not discuss the influence of this on the performance of tested techniques. Jönsson and Eklundh (2002, 2004) briefly discussed the more accurate derivation of start of season (SOS) and end of season (EOS) measures by the asymmetric Gaussian function-fitting over the Savitzky–Golay filter, BISE algorithm, and Fourier transforms, while Beck et al. (2006) demonstrated the advantages of extracting SOS and EOS metrics from NDVI time series to which the double logistic function-fitting had been applied. Although the above do provide...
useful contributions, they are not detailed investigations of the potential effects that particular factors such as noise level or phenological variation may have on the reduction of noise in NDVI time series.

The current literature on NDVI noise-reduction comparisons is presently lacking in two important respects. First, previous efforts have evaluated only a small selection of the noise-reduction techniques currently available. While certainly informative, only one work (van Dijk et al., 1987) has compared more than three techniques concurrently, and the majority of others (e.g., Chen et al., 2004; Ma & Verostraete, 2006; Viovy et al., 1992) have focused on contrasting single innovative methods against one or two widely-known standards: the BISE algorithm and Fourier transforms. In such comparisons, the old standards are usually out-performed by the new techniques, but one is left to wonder: how would these new strategies compare against one another? And how would they perform when applied alongside other demonstrably successful but little known techniques, such as Velleman’s (1980) 4253H, Twice filter? Second, previous evaluations have principally comprised qualitative (i.e., visual inspections) rather than quantitative assessments, and generally do not address factors that could affect the performance of noise-reduction techniques. Qualitative evaluations may only reveal the technique that produces the most visually pleasing result, rather than the most accurate.

This paper aims to address the current need for a more thorough comparison and a broader understanding of the available NDVI time series noise-reduction techniques. The objectives of our research were: (i) to determine which of a selected set of noise-reduction methods performed best under given conditions through an empirical analysis, and (ii) to explore the factors (e.g., level of noise, annual variations, etc.) that might influence the performance of these techniques. A model environment, constructed using actual
remotely-sensed NDVI data over a study area located in west-central Alberta, Canada, was used to address these two objectives. The NDVI time series noise-reduction techniques selected for comparison were two function-fitting techniques and four filters, representing some of the best and most easily-accessible strategies currently available.

2. Methods

In order to establish an analytical framework within which noise-reduction strategies could be evaluated empirically and under a variety of conditions, we undertook our analysis within a model environment comprising a series of ideal modeled NDVI time series representing a variety of biogeographical regions, randomly disturbed with varying levels of introduced noise. Candidate noise-reduction techniques were judged on (i) their capacity to return individual noisy NDVI pixels to their true values, and (ii) their ability to preserve the integrity (amplitude and shape) of the overall time series, so that phenological metrics could be accurately extracted. Fig. 1 outlines the basic steps and procedures followed in our analysis.

2.1. Study area

The study area covers approximately 71,500 km², and is located along the front ranges of the Rocky Mountains in west-central Alberta, Canada (Fig. 2). It represents a highly seasonal environment with a wide diversity of landscapes and ecosystems that can be divided into a number of natural regions and subregions on the basis of common vegetative, physiographic, and soil patterns (Natural Regions Committee, 2006). These include a Grassland region (subregions Mixedgrass and Foothills Fescue), a Parkland region (subregions Central and Foothills Parkland), a Boreal region (subregion Dry Mixedwood), a Foothills region (subregions Lower and Upper Foothills), and a Rock Mountain region (subregions Montane, Subalpine and Alpine). A detailed description of these areas is provided by the Natural Regions Committee (2006).

2.2. Data

For this analysis, we used the MOD13Q1 (version 004) 16-day 250 m NDVI and NDVI Quality Assurance products from Terra's Moderate Resolution Imaging Spectroradiometer (MODIS). These data were acquired from the Earth Observing System Data Gateway on the Land Processes Distributed Active Archive Center website (EOS, 2006). The imagery covered a three-year time series, from January 1, 2003 to December 31, 2005. We used the MODIS Reprojection Tool 3.0 to extract the desired bands and re-project into Universal Transverse Mercator (Zone 11, NAD83).

2.3. Constructing the model environment

A series of six model NDVI time series were constructed, each representing a natural region and/or subregion from the study area. Natural subregions were assumed to represent largely homogenous ecological and vegetative communities demonstrating a distinctive yearly phenological cycle that could be represented for our purposes by a single modeled NDVI time series. We selected a minimum of 100 high-quality pixels for each subregion, identified as those having the highest number of ‘good’ or better vegetation index usefulness index values—a quality index provided with MODIS data sets (LP DAAC, 2005). Date-by-date averaging of these high-quality pixels was used to produce an initial modeled time series for each subregion. Any negative NDVI values were set to zero, based on the assumption that these represented an absence of vegetation and were thus data errors (Myneni et al., 1995). Next, all winter NDVI values in the modeled time series (dates from November to February) were set to an NDVI of 0.1 or
0.15, depending on the subregion. This was based chiefly on visual inspection of the time series – the majority of modeled winter values were between 0.0 and 0.2 – as well as the work of DeFries and Townshend (1994) who demonstrated a temporally-consistent NDVI of 0.1 for bare soil. The winter period was assumed to represent a time during which photosynthesis should be at a minimum, or altogether dormant, at the latitude of the study area. Finally, the sudden transitions between winter dormancy and the growing season were softened by increasing the first and last NDVI values of the winter season by 5% of the NDVI range from dates directly adjacent to them, since it is often recognized that NDVI time series should reflect a smoothly-progressing pattern of vegetation growth and development with time (Ma & Veroustraete, 2006; van Dijk et al., 1987; Wang et al., 2005).

Upon initial model construction, all but two of the subregions – Upper Foothills and Subalpine – produced relatively smooth final modeled time series that were deemed suitable for further analysis. The Upper Foothills and Subalpine time series were noisy, and were dropped from subsequent analysis. In addition, the two modeled Grassland time series – Mixedgrass and Foothills Fescue – were almost identical, and so were averaged to produce a single modeled time series in order to reduce redundancy. The two Parkland subregion modeled time series – Central Parkland and Foothills Parkland – were similarly redundant and were also combined.

In order to synthesize increasingly-noisy time series from the modeled ideals, we introduced noise into each of the six modeled NDVI time series by replacing a random selection of 10%, 40% and 70% of the dates (i.e. low, moderate and high levels of noise) with values from noisy pixels in the study area (Fig. 3). The noisy pixels were selected as those with the greatest number of ‘average’ or worse dates from each appropriate subregion or region, again identified using the vegetation index usefulness index.

### 2.4. Candidate noise-reduction techniques

We selected six candidate noise-reduction techniques for empirical evaluation (Table 1), choosing strategies from the literature that had demonstrated successful application elsewhere, and were likely to be both readily accessible to researchers and suitable for application under a variety of non-specialized conditions. The selected techniques included two function-fitting algorithms: Jönsson and Eklundh’s (2002) asymmetric Gaussian method and Beck et al.’s (2006) double logistic method. Both were successfully applied to multi-temporal NDVI data sets over Africa and northern Europe, respectively, and both are easily implemented using Jönsson and Eklundh’s (2004) TIMESAT program (http://www.natgeo.lu.se/personal/Lars.Eklundh/TIMESAT/timesat.html). Alternative function-fitting procedures, such as Fourier-based techniques, were not considered because of their complexity, their demonstrated inability to deal with irregular or asymmetrical time series, and their potential to cause spurious oscillations in the data (Swets et al., 1999; van Dijk et al., 1987). We also selected four filtering techniques for evaluation: Chen et al.’s (2004) modified Savitzky–Golay filter, Ma and Veroustraete’s (2006) mean-value iteration (MVI) filter, Velleman’s (1980) 4253H, Twice filter, and Filipova-Racheva and Hall-Beyer’s (2000) autoregressive combination ARM3-ARMA5 filter. Each of these has demonstrated superior performance over alternative filters (e.g. BISE), and each can be applied easily to a variety of NDVI time series.

Except for the ARM3-ARMA5 filter, the candidate techniques are described in detail in the literature (Table 1). We employed them as they are described by these authors: we used a 10% multiyear average for the MVI filter threshold, and retained the recommended parameters provided in the TIMESAT program for the two function-fitting methods and the Savitzky–Golay filter (Jönsson & Eklund, 2006). For the latter, we selected window widths of five, six, and seven for the three fitting iterations, based on visual inspections of the data. ARM3-ARMA5 applies two autoregressive moving window filters to an NDVI time series (Filipova-Racheva & Hall-Beyer, 2000): the first is an autoregressive moving median algorithm with a 3-element window, followed by an autoregressive moving mean algorithm with a 5-element window. The simplicity of the ARM3-ARMA5 filter and its demonstrated success in the work of Filipova-Racheva and Hall-Beyer (2000) make it an appealing sixth candidate technique. It should be noted that of the six chosen techniques, all but the 4253H, Twice and ARM3-ARMA5 filters are designed to account for

![Image](image103x622_to_501x741)

**Fig. 3.** The random selection (circles) of 40% of the dates in the Montane modeled NDVI time series (black line), along with the noisy time series (grey line) from which noise was introduced (a), and the resulting Montane modeled NDVI time series with 40% introduced noise (b).

### Table 1

<table>
<thead>
<tr>
<th>Candidate technique</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetrical Gaussian</td>
<td>Fits local, nonlinear functions at intervals around local maxima and minima, then merges these into a global function describing the full NDVI time series</td>
<td>Jönsson and Eklundh (2002)</td>
</tr>
<tr>
<td>function-fitting filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double logistic function-fitting filter</td>
<td>Uses a series of parameters (e.g. winter, maximum NDVI) to model NDVI time series with a double logistic function</td>
<td>Beck et al. (2006)</td>
</tr>
<tr>
<td>Savitzky–Golay filter</td>
<td>Applies an iterative weighted moving average filter to NDVI time series, with weighting given as a polynomial of a particular degree</td>
<td>Chen et al. (2004)</td>
</tr>
<tr>
<td>4253H, twice filter</td>
<td>Applies a series of running medians of varying window size and a weighted average filter, with re-roughing, to the NDVI time series</td>
<td>Velleman (1980)</td>
</tr>
<tr>
<td>Mean-value iteration filter</td>
<td>Iteratively compares each date with the average of the dates before and after it, replacing the date with this average if the difference is above a certain threshold</td>
<td>Ma and Veroustraete (2006)</td>
</tr>
<tr>
<td>ARM3-ARMA5 filter</td>
<td>Applies an autoregressive running median filter with a window size of 3, followed by an autoregressive running mean filter with a window size of five</td>
<td>Filipova-Racheva and Hall-Beyer (2000)</td>
</tr>
</tbody>
</table>
negatively-biased noise in NDVI data, and aim to preserve the upper envelope of NDVI values.

2.5. Analysis

We used two complementary approaches to evaluate the performance of each noise-reduction technique: (i) root mean square error (RMSE) analysis, and (ii) metric analysis. The RMSE analysis (Willmott, 1982) measured the root of the mean squared difference between the modeled NDVI time series (assumed to be truth) and the corresponding experimental time series to which noise reduction had been applied. We interpreted the RMSE results as an evaluation of the overall capacity of a technique to return noisy NDVI time series to their true values. The metric analysis, alternatively, involved the comparison of phenologically-based metrics such as start, end, and length of growing season derived from the modeled NDVI time series (assumed to be truth) to the same metrics derived from the experimental noise-reduced time series (Table 2). We interpreted the results of the metrics analysis as an evaluation of the ability of each candidate technique to maintain the integrity (i.e. amplitude and shape) of the true phenological signal.

2.5.1. Evaluating overall performance

In order to facilitate comparative evaluations of both RMSE and metric analysis, and to provide an overall measure of noise reduction performance, we standardized the raw results into performance scores. Raw metric results were first normalized by dividing each metric estimate from the noisy and noise-reduced time series by the corresponding model-derived metric (i.e. truth) using the general formula:

\[
RM_{\text{syn}} = \frac{M_{\text{syn}}}{M_{\text{MMry}}}
\]

Where \(RM_{\text{syn}}\) is the relative error metric for a particular biogeographical region \(r\), year \(y\) and noise level \(n\), \(M_{\text{MMry}}\) is the metric calculated from a noisy or noise-reduced NDVI time series of the same biogeographical region, year and noise level, and \(M_{\text{MMry}}\) is the original model time series-derived metric for the particular biogeographical region and year. This was necessary, since some metrics are time-based (e.g. start of season), while others are value-based (e.g. maximum NDVI). Normalized metric results were then transformed into unstandardized metric performance scores using a categorization technique wherein an increasing score reflected increasing under- or over-estimation from the model. For instance, a normalized maximum NDVI metric of 0.9 indicated the estimation to be 90% of the model-derived metric, or an unstandardized performance score of 1.

RMSE results did not require normalization, but were simply scaled and rounded into unstandardized RMSE performance scores that matched the scale and range of the unstandardized metrics performance scores in order to facilitate comparison.

The unstandardized performance scores from the metric and RMSE analyses were then summed to produce standardized summary performance scores that provided a measure of overall noise reduction performance. Overall standardized RMSE and metric performance scores were calculated for each candidate technique as

\[
\text{score}_{ij} = \frac{\sum j \text{ score}_i}{n}
\]

Where \(\text{score}_{ij}\) is the standardized score for the \(i\)th noise-reduction technique for a particular category (i.e. a biogeographical region, noise level, year, or metric), \(\text{score}_i\) is the total RMSE or metric unstandardized performance score summation for the \(i\)th noise reduction technique for a particular category, \(\text{score}_i\) is the unstandardized performance scores of the \(n\) noise-reduction techniques for a particular category, and \(n\) is the number of noise-reduction techniques. The results were then summed to form overall standardized summary performance scores for each candidate technique.

In addition to overall standardized summary performance scores, we also derived stratified standardized summary performance scores by calculating summary performance scores on the basis of biogeographical region (e.g., Grassland, Parkland, etc.), metric (e.g., SOS, EOS, etc.), year (e.g., 2003, 2004, 2005) and noise level (e.g., 10%, 40%, 70%). The summary and stratified standardized performance scores provided relative measures of candidate technique performance with regard to the other strategies under evaluation, including the ‘do nothing’ strategy of applying no noise reduction to the experimental time series. It is important to note that because they have been normalized (in the case of metric results) and standardized these performance scores are not ratio-level measures of evaluation, but rather provide a relative means of assessing each technique comparatively. That is, a standardized performance score of 1.0 does not indicate a performance that is 100% better than a score of 2.0, but rather, simply denotes a notably superior performance.

3. Results

3.1. Overall performance

Fig. 4a shows the overall standardized summary performance scores for the six candidate noise-reduction techniques. The two function-
fitting techniques – asymmetric Gaussian and double logistic – performed the best overall with standardized summary performance scores of 1.79 and 1.78, respectively. All six noise-reduction techniques produced better (i.e., lower) overall performance scores (1.78 to 2.10) than the noisy NDVI time series to which no noise reduction had been applied (2.48). However, more complex patterns emerged when the standardized RMSE and metric summary performance scores were compared separately (Fig. 4b). For example, the Savitzky–Golay filter (0.87) out-performed all other noise-reduction techniques when evaluated on the basis of RMSE, but ranked worse (1.06) than three other techniques – asymmetric Gaussian (0.89) and double logistic (0.88) function-fitting, and the 4253H, Twice (0.88) filter – when

---

Fig. 4. Overall summary standardized (a) and RMSE and metric standardized (b) performance scores for the six candidate noise-reduction techniques and no noise reduction (‘None’). Technique abbreviations for all figures are as follows: double logistic function-fitting (DL), asymmetric Gaussian function-fitting (AG), Savitzky–Golay filter (SG), and mean-value iteration filter (MVI).

Fig. 5. Standardized summary performance scores for the six candidate noise-reduction techniques and no noise reduction (‘None’), stratified by biogeographical region (a), level of introduced noise (b), year (c), and metric (d).
evaluated on basis of metrics performance scores. Thus, while the Savitzky–Golay filter was better able to minimize overall noise in NDVI time series, the two function-fitting procedures and the 4253H, Twice filter proved better able to maintain the integrity of the time series for the purpose of extracting metrics.

3.2. Stratified performance

Fig. 5a illustrates the standardized summary performance scores stratified by biogeographical region. Considerable variation in performance was observed, but two patterns can be discerned. First, over all but the Parkland and Montane landscapes, the six candidate noise-reduction techniques each performed better than the raw NDVI time series to which no noise reduction had been applied. Second, the two function-fitting techniques generally performed better overall with standardized performance scores often between 1.65 and 1.75, while the MVI and ARMD3-ARMA5 filters generally performed worse, often showing scores between 2.0 and 2.3 (Table 3).

Fig. 5b shows the results of standardized performance scores stratified by noise level, where three principal observations can be made. First, no one noise-reduction technique out-performed all others across the three noise levels we investigated; the performance of each technique varied with noise level. Second, at low noise levels (10%), only two techniques (the double logistic function-fitting and 4253H, Twice filter) performed better than the raw time series to which no noise reduction had been applied. In other words, from an overall standardized performance score perspective, four of the candidate noise-reduction techniques tended to make things worse when noise levels were low. However, the trend did not exist at moderate (40%) or high (70%) noise levels, where the raw, unprocessed time series produced the worst overall scores at 2.95 and 2.63, respectively (Table 3).

Fig. 5c shows the stratification of standardized metric performance scores by year. The results indicated that while some annual variation did exist, performance was generally consistent from year to year (Table 3). The double logistic and asymmetric Gaussian function-fitting procedures and the 4253H, Twice filter all produced consistently better scores than the three alternative filtering techniques across the three years under investigation. Greater variation existed between the performances of the six candidate techniques in any given year than for one particular technique across the three years.

Finally, Fig. 5d shows the standardized metric performance scores as stratified by metric. These results showed large amounts of variation, revealing the notable sensitivity of noise reduction performance to individual metric. For example, the Savitzky–Golay filter produced both the best (maximum NDVI) and worst (maximum green-up) scores recorded in our analysis, at 0.53 and 1.75, respectively (Table 3). Despite this substantial variability, two general observations were made. First, both function-fitting procedures outperformed the MVI and ARMD3-ARMA5 filters for all metrics but maximum NDVI, while neither the 4253H, Twice filter nor the Savitzky–Golay filter were consistent in their performance with metric. Second, although the original noisy time series did not always produce the worst scores, four of the five time-based metrics – SOS, EOS, LGS and the timing of maximum green-up – were consistently improved with the application of noise reduction. The extraction of other metrics, particularly non time-based measures such as maximum NDVI, maximum rate of green-up and integrated-NDVI were not consistently improved with noise reduction.

4. Discussion

4.1. General trends

In our experiments, the asymmetric Gaussian and double logistic function-fitting techniques performed best when evaluated on the basis of overall noise reduction (i.e. RMSE analysis) and of data integrity (i.e. metric analysis) over a variety of conditions reflecting those found in the western Alberta study area. These findings support the conclusions drawn by Jönsson and Eklundh (2002, 2004) and Beck et al. (2006). The similarity in the performance of the two function-fitting techniques is not surprising, given the strong resemblance in their respective implementations, and is likely the result of two factors: (i) strict preservation of the upper envelope of NDVI values, and (ii) reliable approximation of constant winter NDVI, reflecting that season’s dormant vegetative condition at mid- to high latitudes. This is illustrated in Fig. 6, where the asymmetric Gaussian and double logistic function-fitting techniques were better able to approximate winter conditions and minimize negatively-biased noise than the alternatives. Both showed effective maintenance of higher NDVI values during the growing season, while the MVI and ARMD3-ARMA5 filters tended to underestimate these values noticeably, the 4253H–Twice filter underestimated these values slightly, and the Savitzky–Golay filter tended to overestimate them. In addition, neither the asymmetric Gaussian nor the double logistic technique was affected by the sudden dips in composite periods at the beginning and end of 2003, and at the end of 2005, as were the Savitzky–Golay filter and to some degree, the 4253H–Twice filter. Both the MVI and ARMD3-ARMA5 filters showed similar winter NDVI trajectories, but neither was able to approximate the suddenness of the beginning or end of the three growing seasons as were the two function-fitting methods.

Both the MVI and ARMD3-ARMA5 filters showed generally poor performance, both in overall evaluations, and in the RMSE and metric-based evaluations. These findings contradict those of Ma and Veroustraete (2006) and Filipova-Racheva and Hall-Beyer (2000), though it must be recognized that these previous evaluations were primarily visual and qualitative in nature. In our experiments, the shared reliance of MVI and ARMD3-ARMA5 filters on running means, combined with their inability to preserve the upper envelope of NDVI values, resulted in poor performances overall. As both van Dijk et al. (1987) and Jönsson and Eklundh (2006) observed, even weighted averaging techniques tend to alter the shape and amplitude of an NDVI

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Standardized summary performance scores for the six noise-reduction techniques and no noise reduction ('None'), stratified as shown in Fig. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biogeo. region</td>
<td>4253H, Twice</td>
</tr>
<tr>
<td>Grassland</td>
<td>1.87</td>
</tr>
<tr>
<td>Parkland</td>
<td>1.73</td>
</tr>
<tr>
<td>Boreal</td>
<td>1.74</td>
</tr>
<tr>
<td>Lower Foothills</td>
<td>2.04</td>
</tr>
<tr>
<td>Montane</td>
<td>1.52</td>
</tr>
<tr>
<td>Alpine</td>
<td>2.08</td>
</tr>
<tr>
<td>Noise level</td>
<td>10% Noise</td>
</tr>
<tr>
<td>40% Noise</td>
<td>1.76</td>
</tr>
<tr>
<td>70% Noise</td>
<td>1.99</td>
</tr>
<tr>
<td>Year</td>
<td>2003</td>
</tr>
<tr>
<td>2004</td>
<td>0.84</td>
</tr>
<tr>
<td>2005</td>
<td>0.90</td>
</tr>
<tr>
<td>Metric</td>
<td>Max NDVI</td>
</tr>
<tr>
<td>(r) Max NDVI</td>
<td>0.75</td>
</tr>
<tr>
<td>NDVI amplitude</td>
<td>0.94</td>
</tr>
<tr>
<td>SOS</td>
<td>0.86</td>
</tr>
<tr>
<td>EOS</td>
<td>0.75</td>
</tr>
<tr>
<td>LCS</td>
<td>0.83</td>
</tr>
<tr>
<td>Max green-up</td>
<td>0.79</td>
</tr>
<tr>
<td>(r) Max green-up</td>
<td>0.99</td>
</tr>
<tr>
<td>I-NDVI</td>
<td>0.95</td>
</tr>
<tr>
<td>Average NDVI</td>
<td>0.97</td>
</tr>
</tbody>
</table>
time series, and can be influenced significantly by large outliers (Velleman 1980). Thus, rapid spring green-up events or sharp autumnal senescence are more likely to be distorted from the original time series by either of these filters. In addition, since neither method accounts for the common negative bias in NDVI noise caused by atmospheric and snow effects (Gutman, 1991), they were less capable of dealing with such noise than alternative techniques.

While both the Savitzky–Golay filter and 4253H, Twice filter performed well under certain conditions in our analysis, neither was consistent in general performance across the three sets of standardized summary performance scores. The greater ability of the Savitzky–Golay filter to minimize overall noise in NDVI time series is likely linked with its tendency to preserve higher NDVI values and account for negatively-biased noise. Indeed, these findings support the successful application of this filter demonstrated by Chen et al. (2004). However, the inferior ability of this filter to maintain original signal integrity may relate to the sensitivity of the technique to window widths used in the technique’s fitting procedure. That is, a wide window limits the ability of the filter to follow rapid but relevant changes in NDVI (e.g. a rapid green-up), while a narrow window may cause the result to over-fit the time series and retain more noise (Chen et al., 2004; Jönsson & Eklundh, 2006). This sensitivity renders the filter less reliable under varying conditions. Unlike the Savitzky–Golay filter, however, the 4253H, Twice filter does not preserve the upper envelope of NDVI values, which doubtless accounted for its poorer performance in general noise minimization. Nonetheless, the reliance of this filter on running medians likely accounted for its greater ability to maintain original signal integrity, since running medians are able to remove sharp discontinuities in a time series while preserving the quadratic shape of a clean NDVI profile (Rabiner et al., 1975). In fact, the ability of the 4253H, Twice filter to handle complex time series while eliminating spurious drops and spikes was praised by both Velleman (1980) and van Dijk et al. (1987).

Finally, the general benefit of applying any of the six candidate noise-reduction techniques was clear. The noisy NDVI time series performed consistently poor in all three sets of standardized summary performance scores, suggesting that any attempt at rectification was generally better than none at all. Nonetheless, stratification revealed that under particular conditions, such as minimal levels of noise or the subsequent extraction of particular metrics, the use of time series noise reduction may not be ideal.

4.2. Factors influencing performance

The choice of phenologically-based NDVI time series metric was the most influential factor affecting the performance of any given
noise-reduction technique. Performance varied considerably with metric, though any of the six candidate noise-reduction techniques proved beneficial for estimating most time-based metrics (i.e. SOS, EOS, LGS, and time of maximum green-up). High-frequency fluctuations in a noisy NDVI time series are easily mistaken for the timing of real phenological events. Fig. 7 illustrates this, showing cases where noise reduction improved the prediction of SOS and EOS events. These same high-frequency fluctuations can also lead to imprecise estimations in the timing of both maximum NDVI and maximum rate of green-up from noisy time series. On the basis of these observations, we would recommend that if the derivation of metrics related to the timing of the growing season and vegetative development is of particular interest, noise reduction should be considered. However, caution is advised when considering metrics unrelated to phenological timing, such as maximum NDVI, maximum rate of green-up or integrated-NDVI, where the benefit of noise reduction is less evident.

The second factor to demonstrate significant influence on performance was biogeographical region. Although the asymmetric Gaussian and double logistic methods consistently out-performed both the MVI and ARMD3-ARMA5 filters, and across the majority of regions the noisy data produced the worst scores, considerable variation in performance was observed. One would expect this variation to relate to phenological differences, but closer examination revealed the majority of this influence to be caused by differences in noise patterns introduced into each modeled time series. That is, two noisy time series with equivalent levels of introduced noise could show very different patterns in terms of the strength and nature of this noise, and different noise-reduction strategies handled this noise differently. For example, in Fig. 8 the double logistic function-fitting was better able to cope with the negatively-biased noise introduced to the Grassland modeled NDVI time series, but was out-performed by the 4253H, Twice filter over the Montane time series where the noise did not show such a bias. A similar pattern was observed for the Parkland time series with equivalent noise, but in this case the strength of the noise was considerably less, and each technique was better able to handle it. This reflects the observation by Jónsson and Eklundh (2002) that the Savitzky–Golay filter could be problematic when applied to very noisy time series. Based on these findings, we recommend that the strength and character of the noise present in an NDVI data set be considered when selecting an approach to time series noise reduction.

Neither the level of introduced noise nor the year was as influential on noise-reduction technique performance as the other factors discussed above. No one technique performed best over the three levels of introduced noise, though the results suggest that applying noise reduction to relatively clean time series will tend to degrade it. Annual variation in performance was slight. Year did not show any considerable influence on performance, and what annual variation did exist was likely the result of differences in the noise present in each of the three years.

While the above discussion offers valuable insight, we recognize two important limitations. First, it must be noted that our research concerns continental, mid-latitude biomes characterized by a single, distinct annual growing season (e.g. temperate forest and grassland). Our findings and conclusions may not be as applicable to biomes exhibiting multiple growing seasons (e.g. subtropical zones), those displaying little annual seasonality (e.g. the tropics or desert shrubland), or agricultural regions. Second, our results relate specifically to the NDVI. While similar ratio-based indices such as the simple ratio (Jordan, 1969) may produce similar results, alternative vegetation indices such as the soil-adjusted vegetation indices (Huete, 1988; Qi et al., 1994), atmospherically resistant vegetation indices (Kaufman and Tanré, 1992) or the enhanced vegetation index (EVI; Huete et al., 1999) may not. For instance, the increasingly popular EVI is less influenced by atmospheric and background effects than the NDVI, and, unlike the NDVI, is spuriously amplified by the presence of snow. The EVI thus tends to have less negatively-biased noise and more erroneous spikes than the NDVI, in which case noise-reduction techniques maintaining the upper envelope

![Fig. 7. Boreal (a, b), and Lower Foothills (c, d) modeled, noisy and noise-reduced NDVI time series, showing 40% (left) and 70% (right) introduced noise, and application of the 4253H, Twice filter. Start of growing season dates (circles) and end of growing season dates (squares) are also shown.](image-url)
of values such as the double logistic and asymmetric Gaussian function-fitting techniques may not be the most effective choice.

5. Conclusion

By performing an empirically-based comparison of several selected methods, we have demonstrated the general superiority of Jönsson and Eklundh’s (2002) asymmetric Gaussian and Beck et al.’s (2006) double logistic function-fitting techniques over four alternative filters. Both showed a balanced ability to reduce noise while maintaining the relevant NDVI signal integrity. However, stratification revealed the considerable influence of metric choice and biogeographical region on the performances of all six candidate techniques. Not only did the strength and character of the noise present in each biogeographical region influence results, but so did the choice of phenologically-based metric. Annual variations and the impact of noise level were minimal. On the basis of these findings we offer two recommendations. First, the strength and nature of the noise present in a data set should be considered when selecting a method for NDVI time series noise reduction. Where no distinct negative bias is present, the 4253H, Twice filter performs well, while either of the two function-fitting methods would be more properly applied in situations of negatively-biased noise. Second, we advise caution in the application of such techniques when the extraction of time series metrics is the ultimate aim, as some time-based metrics such as SOS may benefit from noise reduction while other value-based metrics may not.

Noise reduction in NDVI time series is neither simple nor straightforward. However, our use of a controlled model environment, an approach that is rarely if at all found in the literature, allowed for a more exhaustive exploration of the selected techniques than would otherwise have been possible. Of primary interest to future work is the verification of the present results within a real-world analysis. This would prove highly valuable, were the costs and difficulties of acquiring the required ground truth data minimized, perhaps through the use of proxy data. Flux tower measurements collected through the FLUXNET project might offer such a proxy. Another crucial avenue for further investigation concerns a greater understanding of the effects of NDVI time series noise reduction on the subsequent extraction of phenologically-based metrics. The present study reveals the potential inaccuracies that may result when deriving metrics from noise-reduced time series, but does not rigorously explore this relationship. Further work is needed if the implications of applying noise reduction in the extraction of these metrics are to be clearly understood.

Acknowledgement

We would like to thank Dr. Myrka Hall-Beyer for her insightful assistance. This research was generously supported by the Natural
Sciences and Engineering Research Council of Canada and the Alberta Ingenuity Fund. Comments from three anonymous reviewers helped improve the final version of the manuscript.

References


