Available online at www.sciencedirect.com
ScienceDirect
Engineering Applications of Artificial Intelligence 19 (2006) 787–795
Engineering Applications of
ARTIFICIAL
INTELLIGENCE
www.elsevier.com/locate/engappai

A statistical complement to deterministic algorithms for the retrieval of aerosol optical thickness from radiance data

Bo Han[a], Slobodan Vucetic[a], Amy Braverman[b], Zoran Obradovic[a,*]

[a]Center for Information Science and Technology, Temple University, 1805 N. Broad St, Philadelphia, PA 19122, USA
[b]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

Received 18 May 2006; accepted 21 May 2006
Available online 14 August 2006

Abstract

As a complement to the conventional deterministic geophysical algorithms, we consider a faster, but less accurate approach: training regression models to predict aerosol optical thickness (AOT) from radiance data. In our study, neural networks trained on a global data set are employed as a global retrieval method. Inverse distance spatial interpolation and region-specific neural networks trained on restricted, localized areas provide local models. We then develop two integrated statistical methods: local error correction of global retrievals and an optimal weighted average of global and local components. The algorithms are evaluated on the problem of deriving AOT from raw radiances observed by the Multi-angle Imaging SpectroRadiometer (MISR) instrument onboard NASA's Terra satellite. Integrated statistical approaches were clearly superior to global and local models alone. The best compromise between speed and accuracy was obtained through the weighted averaging of global neural networks and spatial interpolation. The results show that, while much faster, statistical retrievals can be quite comparable in accuracy to the far more computationally demanding deterministic methods. Differences in quality vary with season and model complexity.

Keywords: Geophysical retrievals; Aerosols; Regression

1. Introduction

NASA's Earth Observing System (EOS) satellites, Terra, Aqua, and Aura, have now been in orbit for several years. The instruments aboard these platforms provide steady and massive streams of information on many different geophysical characteristics of the Earth's atmosphere and environment. Their primary purpose is to characterize how the Earth's system is changing, and to identify and understand the primary causes of that variability (Herring and King, 2000).

As a basic principle of remote sensing, satellite instruments measure radiances emitted or reflected from Earth. These data are used to estimate underlying geophysical characteristics, such as atmospheric temperature profiles, cloud and aerosol properties, and the extent of snow, ice, or vegetation cover (Diner and Davies, 2003). The process of inferring these characteristics from observed radiances is called retrieval. The retrieved quantities, called parameters, are then used in various applications ranging from natural resource monitoring to the development of general circulation models for a climate. Accurate and timely retrievals are critical for the success of the subsequent analyses.

One of the most important tasks for the EOS mission teams is the retrieval of aerosol information. Aerosols are small particles produced by natural and man-made sources that both reflect and absorb incoming solar radiation. Aerosol concentration and chemical properties are important parameters in climate change models, in studies on regional radiation balances, and of the hydrological cycle (Ramanathan et al., 2002). Using radiance observations from satellites, it is possible to estimate the attenuation of solar energy as it passes through a column of atmosphere, a quantity commonly known as aerosol optical thickness (AOT).

Since 2000, the Multi-angle Imaging SpectroRadiometer (MISR) instrument has been providing aerosol

*Corresponding author. Tel.: +1 215 2046265; fax: +1 215 2045082.
E-mail address: zoran@ist.temple.edu (Z. Obradovic).

0952-1976/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2006.05.009

information. MISR AOT retrievals provide information about both AOT (also called aerosol optical depth, AOD), and aerosol particle properties. The operational MISR retrieval algorithm is based on separately developed ocean and land forward models, and uses a deterministic inversion approach (Kahn et al., 1997; Martonchik et al., 1998, 2002). The forward MISR model is applied on 24 representative aerosol mixtures to obtain a look-up table with simulated observed radiances. The simulated data are then compared to actual observations for the appropriate scene type (land or ocean), and the nearest neighbor in the look-up table is "retrieved". We call this *the deterministic retrieval*.

Due to the large volume of data being collected by the MISR, compromises are required between retrieval accuracy and processing complexity. For example, the MISR collects raw data at a 1.1 km resolution, but retrieves aerosol properties at 17.6 km resolution for just 24 postulated aerosol types. Processing 650 MB of raw radiance data collected every day requires 30–60 h on a dual processor workstation with a maximum resident memory of 1 GB. Scientifically, it would be preferable to retrieve aerosol properties at 1.1 km resolution and with a much larger number of postulated aerosol types, but this would require an increase in processing resources of several orders of magnitude. The major objective of this study was to explore whether statistical approaches can be used to alleviate this computational hindrance.

Our goal here was to show that statistical approaches could complement deterministic retrieval algorithms to significantly reduce computational costs while introducing only a slight accuracy overhead. The underlying idea was to use deterministic algorithms over a reduced set of locations, and to rely on statistical approaches to provide accurate retrieval over the remaining locations.

We explored two basic statistical approaches. One is based on spatial interpolation and another on the construction of neural network predictors. They both have been successfully used for geophysical parameter retrievals. Zhao et al. (2005) applied spatial interpolation to improve Moderate-Resolution Imaging Spectroradiometer primary vegetation product MOD17. Faure et al. (2002) have shown that the neural network approach could be used to retrieve cloud parameters of inhomogeneous clouds from high-resolution multispectral radiometric data. Berdnik and Loiko (2006) investigated the problem of retrieval of size and refractive index of spherical particles by using multilevel neural networks.

In this study, spatial interpolation utilizes the property that AOT values are strongly spatially correlated over the distances of up to 100 km. In the second approach, neural networks use AOTs provided by deterministic retrievals as training data to directly learn mappings from raw radiance observation attributes to AOT. Both statistical approaches are less accurate than the deterministic algorithm because they directly depend on its retrievals. Since spatial interpolation and neural network approaches are quite different, we also explored algorithms that combine their strengths. To this end, we took two approaches: (1) using spatial interpolation to correct errors in neural network predictors and (2) using weighted averages of retrievals produced separately by spatial interpolation and neural networks.

Our preliminary results (Han et al., 2005) indicated that statistical approaches have interesting trade-offs between retrieval speed and accuracy. This paper reports on a more detailed study by evaluating the preliminary and alternative statistical approaches over substantially larger data sets. The proposed methods were evaluated using MISR data over the continental United States during four 16-day periods in 2002: July 1 to 16, 2002, July 7 to August 1, 2002, October 1to 16, 2002, and October 17 to November 1, 2002. Experimental results provide further indication that statistical approaches represent a promising mechanism for improving aerosol retrieval efficiency, without significantly reducing accuracy.

2. Methodology

2.1. Preliminaries

Given a set $\{x_i\}$ of satellite-based radiance observations, each data point x_i is represented as tuple $x_i = [t_i, \mathrm{lat}_i, \mathrm{lon}_i, o_{i1} \ldots o_{iM}, a_{i1} \ldots a_{iN}]$, where t_i is the time of the observation, lat_i and lon_i denote the spatial location, o_{ij}, $j = 1 \ldots M$, are attributes derived from the observed radiances, and a_{ij}, $j = 1 \ldots N$, are ancillary attributes. For example, geometric parameters describe instrument camera view angles and sun angle. The aerosol retrieval task can be stated as a prediction of AOT, denoted as y_i, based on the corresponding values of x_i.

Deterministic retrievals are based on a forward model $o_i = F(a_i, y_i)$, which estimates what the observed radiances o_i will be based on the aerosol properties y_i and ancillary attributes a_i. The forward model is derived directly from physical principles. Given the forward model F, the remaining task is to derive a retrieval algorithm as an inverse operator of F, $R = F^{-1}$. The development of a retrieval algorithm by the inversion of the nonlinear forward model is an ill-posed problem. The deterministic retrieval is based on the construction of a look-up table that consists of a large number of tuples $(a_i, y_i, F(a_i, y_i))$ corresponding to a number of retrieval scenarios defined by (a_i, y_i) tuples. Then, given an observation defined by pair (a, o), the look-up table entry with values of $(a_i, F(a_i, y_i))$ closest to the observation is found. The value y_i of this entry is reported. This inversion procedure is equivalent to nearest neighbor algorithms and is computationally costly because it requires linear search time for each retrieval.

Statistical retrievals use a labeled data set with tuples $\{(x_i, y_i), i = 1 \ldots K\}$ to provide retrievals at desired locations. In the case of neural networks, statistical retrieval algorithms learn an accurate regression model $\mathrm{NN}(x, \beta)$ that minimizes the Mean Squared Error (MSE) defined as

$E[(y - \mathrm{NN}(x, \beta))^2]$, where β is a set of model parameters. While learning regression models can be time consuming, the retrieval (i.e., evaluation of $\mathrm{NN}(x, \beta)$ for a given x) is rapid. Alternatively, the spatial correlation of aerosols can be used to predict AOT at locations in the vicinity of points represented by the labeled data set.

2.2. Problem definition

Based on the properties of deterministic and statistical retrieval algorithms described in Section 2.1, it is evident that the two approaches are complementary—deterministic retrieval algorithms are accurate and computationally expensive, while statistical algorithms may be less accurate and are significantly less expensive. The goal of this paper is to explore how to combine the two approaches to achieve a trade-off between accuracy and computational complexity. The approach explored in our study is as follows:

- Instead of applying the deterministic algorithm at high spatial resolution, we apply it over a significantly reduced number of spatial locations and use a statistical algorithm to provide retrievals at the remaining locations.

The objective here is to explore the potential of this approach to significantly reduce retrieval time, with only a slight decrease in accuracy. In Sections 2.3–2.7 we outline five different statistical retrieval algorithms to be combined with the deterministic algorithm used by MISR.

2.3. Retrieval by Global Artificial Neural Networks (GANN)

Assume we are given a set of K labeled data points $\{(x_i, y_i), i = 1 \ldots K\}$ where labels y_j are obtained from a deterministic algorithm over some period of time. We construct a GANN regression model $y_i = \mathrm{GANN}(x_i)$ by using non-spatial attributes x_i (without including $(\mathrm{lat}_i, \mathrm{lon}_i)$) as input, and targets y_j as output. The trained model is called global because it uses previously retrieved data in all regions for its training.

In our experiments, GANN was structured as a feedforward neural network, with a single hidden layer having sigmoid neurons and a linear neuron at the output. The design objectives were dimension reduction and identification of an optimal GANN structure in order to maximize the prediction accuracy on separate test data. Prediction accuracy is measured by MSE, or by the coefficient of determination $R^2 = 1 - \mathrm{MSE}/\mathrm{Var}(y)$, where $\mathrm{Var}(y)$ is the variance of the AOT target variable.

Publicly available MISR data products allow derivation of a large number of attributes. Direct use of these attributes as inputs would result in a neural network with a large number of weights, a time-consuming training procedure, and possible overfitting problems. Therefore, we applied principal component analysis (PCA) to reduce data dimensionality by using the largest principal components as inputs to our ANN. We determined the appropriate number of hidden units in the ANN by empirically comparing overall prediction accuracies. Based on results of our preliminary study (Han et al., 2005), we selected Bayesian regularization as the ANN training algorithm.

2.4. Retrieval by Inverse Distance Spatial Interpolation (IDSI)

Given a set of K spatially co-located data points $\{(\mathrm{lat}_i, \mathrm{lon}_i, y_i), i = 1 \ldots K\}$ obtained from the deterministic algorithm, IDSI can be used to predict AOT at neighboring locations. The AOT value y at desired location $(\mathrm{lat}, \mathrm{lon})$ is computed as

$$y = \frac{\Sigma_{i=1}^{K} y_i / d_i^p}{\Sigma_{i=1}^{K} 1/d_i^p}, \tag{1}$$

where d_i is the distance between $(\mathrm{lat}_i, \mathrm{lon}_i)$ and $(\mathrm{lat}, \mathrm{lon})$, and p is a parameter determining the extent of emphasis on the nearest neighbors (Vucetic et al., 2000). Larger p values assign larger weights to the near neighbors, where $p = 0$ corresponds to simple averaging, and $p = \infty$ is equivalent to the nearest neighbor predictor. We explored several choices for the value of p to determine the best one for use in all subsequent experiments.

2.5. Retrieval by Region-Specific Neural Networks (RSNN)

More complex interpolation methods are possible. For example, we explored RSNN interpolators. For each spatial region covered by one pass of the instrument, called an orbit (see Fig. 1), a local neural network was constructed using a subset of labeled data corresponding to that orbit. In contrast to GANN, RSNN uses smaller number of neurons because it is trained using a small amount of data collected from a specific region. In addition, the input to RSNN is composed of all available attributes including latitude and longitude. Thereby, RSNN can reflect specific properties of a region by explicitly incorporating information from a spatially constrained region. Using spatial coordinates as attributes allows RSNN to perform spatial interpolation, which is desirable for this type of application.

2.6. Retrieval by Error Correction Models (ECM)

We have observed that retrieval errors from GANN are spatially correlated (see Section 3.2.1). Therefore, if GANN is overestimating at given location, it is highly likely that it will overestimate at neighboring locations (up to about 100 km away). This phenomenon can be exploited by spatial autoregressive modeling. In this study, we used a simple approach: given a set of K_r labeled data points $\{(x_i, y_i), i = 1 \ldots K_r\}$ corresponding to orbit r, $r = 1 \ldots R$,

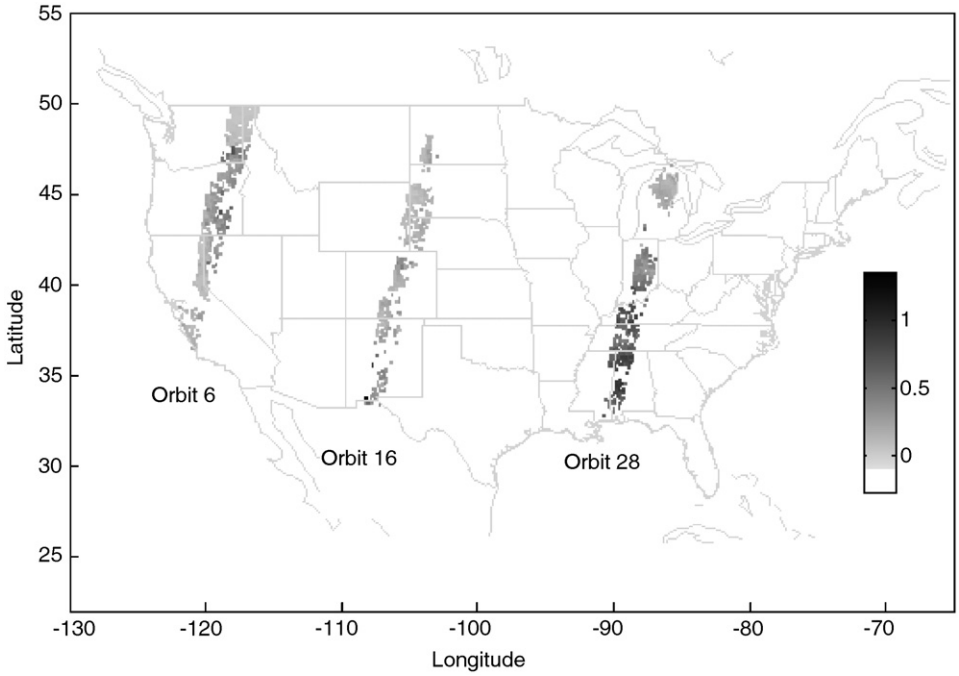

Fig. 1. Examples of three orbits over the continental United States collected on July 15, 2002 (Orbit 6), July 9, 2002 (Orbit 16), and July 5, 2002 (Orbit 28). Missing areas correspond to blocks without MISR retrieval. The intensity represents AOT.

AOT at a location defined by vector x is estimated as

$$\mathrm{ECM}r(x) = \mathrm{GANN}(x) + \varepsilon_r(x), \tag{2}$$

where $\varepsilon_r(x)$ is obtained by IDSI (Eq. (1)) applied to prediction errors $y_i - \mathrm{GANN}(x_i)$ from GANN over the labeled data. In these experiments, the same value of parameter p was used as for original IDSI algorithm of Section 2.4.

2.7. Retrieval by a Weighted Average Model (WAM)

To combine predictions from GANN models and spatial interpolation models, we construct a WAM predictor for each orbit r, $r = 1 \ldots R$, by weighting the component predictors as

$$y = \alpha_r \mathrm{GANN}(x) + (1 - \alpha_r)\mathrm{IDSI}_r(x). \tag{3}$$

Here α_r is the weight parameter constrained to lie in the interval $[0, 1]$.

The weight parameter α_r in Eq. (3) is optimized to minimize the mean-squared prediction error. It can be shown that the optimal choice of α_r is

$$\alpha_r = \frac{\sum_{i=1}^{K_r}(y_i - \mathrm{IDSI}_r(x_i))(\mathrm{GANN}(x_i) - \mathrm{IDSI}_r(x_i))}{\sum_{i=1}^{K_r}(\mathrm{GANN}(x_i) - \mathrm{IDSI}_r(x_i))^2}, \tag{4}$$

where K_r represents the number of labeled data points within the rth orbit. If $\alpha_r < 0$, its value is set to zero; and if $\alpha_r > 1$, its value is set to one.

3. Experimental results

3.1. Data set

MISR data used in our experiments were obtained from NASA's Langley Atmospheric Sciences Data Center. MISR measures reflected solar radiation from nine view angles along the direction of flight and in four spectral bands at each angle (Bothwell et al., 2002; Bull et al., 2005; Mcguckin et al., 1995). On each orbit, MISR sweeps out a 360 km wide swath of data from north to south at a 1.1 km spatial resolution while in daylight. An example of retrieved AOT based on observations during July 2002 is shown in Fig. 1. Since MISR does not collect data at night, consecutive swaths are separated geographically resulting in 14 or 15 evenly spaced half-orbits per day. MISR's ground footprint repeats in 16-day cycles, which is the time it takes the Terra satellite to fly 233 distinct orbital paths (Diner, 1999). We obtained about 66 GB of MISR aerosol, radiance and cloud mask data covering four 16-day repeat cycles over the entire continental US (cycle 1: July 1 to 16, 2002, cycle 2: July 17 to August 1, 2002, cycle 3: October 1 to 16, 2002, and cycle 4: October 17 to November 1, 2002).

The learning target is AOT at 470 nm retrieved at 17.6 km × 17.6 km spatial resolution (EOS Data Gateway, 2005). The attributes listed in Table 1 are derived from 116 radiance variables and geometric parameters (e.g., sun angle, view angle, etc.) measured at 1.1 km spatial resolution and aggregated over each 17.6 km × 17.6 km block. In the preprocessing stage, we used the cloud mask to filter out pixels where at least one of the nine cameras

Table 1
Driving attributes constructed at 17.6 km × 17.6 km resolution from 1.1 km pixels

Attribute index	Name and explanation
1, 2, 3, 4	Time, latitude, longitude, orbit number
5–40	36 mean values of radiance measurements
41–76	36 minimum radiance measurements in each 17.6 × 17.6 km region
77	Solar zenith angle
78–86	View zenith angle (9 cameras)
87–95	Relative view-Sun azimuth angle (9 cameras)
96–104	Scattering angle (9 cameras)
105–113	Glitter angle (9 cameras)
114	Number of cloud-free points in each 17.6 × 17.6 km region
115	Surface feature type in each 17.6 × 17.6 km region

observed clouds. Additionally, the MISR quality flag was used to remove 1.1 km pixels with invalid radiance information. After data cleaning, we obtained 12,304, 10,778, 6616, and 3835 labeled data points for evaluation of the proposed models.

Aerosol properties vary considerably over seasons, and this should be accounted for when interpreting experimental results. Table 2 shows average AOT and its standard deviation over each of the four cycles. AOT is significantly higher and more variable during the two July cycles than during the two October cycles.

3.2. Optimization of statistical retrieval models

In the first set of experiments explained in Sections 3.2.1–3.2.3 we optimize GANN, IDSI, and RSNN models by assuming that 10% of the deterministic AOT retrievals are available for training.

3.2.1. Optimization of GANN

We examined accuracies of GANN with 5, 15, and 30 hidden nodes. We also explored cases when 31 principal components (retaining 95% of attribute variance) and 54 principal components (retaining 99% of the variance) were used as inputs to GANN. Table 3 summarizes the achieved R^2 accuracy for each cycle and every combination of hidden nodes and principle components used to reduce dimensionality. Considering learning complexity and achieved accuracy, GANN with 15 hidden nodes and 31 principal components seemed to be an appropriate choice, and was used in the remaining experiments.

By analyzing the prediction errors of GANN, we observed that they are spatially correlated. As an illustration, in Fig. 2, we show MISR AOT, GANN, and MISR AOT–GANN values for a portion of Orbit 7 on July 10, 2002. Clearly, all three quantities are strongly spatially correlated.

Table 2
Mean and standard deviation of AOT in each cycle

Cycle	AOT mean	AOT standard deviation
Cycle 1: July 1 to 16, 2002	0.268	0.224
Cycle 2: July 17 to August 1, 2002	0.196	0.127
Cycle 3: October 1 to 16, 2002	0.087	0.052
Cycle 4: October 17 to November 1, 2002	0.085	0.047

3.2.2. Optimization of IDSI

To explore the influence of various ways of weighting spatial interpolation towards near neighbors, we ran experiments using different p-parameter values in inverse distance interpolation (Eq. (1)). Results listed in Table 4 suggest that using $p = 2$ is the optimal choice.

3.2.3. Optimization of RSNN

The RSNN should be constructed using a small amount of regional training data. Based on our pilot study (results not shown), an appropriate choice appears to be the five largest principal components together with latitude and longitude attributes. A minimum of 50 training points per orbit were required for RSNN implementation. Thus, this method could be used only for highly populated orbits. Usually, such orbits have few clouds.

3.3. Comparison of statistical algorithms by using 10% of deterministic AOT retrievals for training

Results comparing overall accuracies of GANN, RSNN, IDSI, ECM, and WAM are listed in Table 5. WAM performed best in all cycles. ECM significantly outperformed GANN, IDSI, and RSNN during three cycles. We also observed that performance of the five statistical models can vary considerably over different orbits. To analyze the models' R^2 accuracy in more detail, we list the paired model comparisons in Table 6. Results show the number of orbits (with more than 100 labeled points) for which the first model was more accurate than the second model. WAM performed better than the others on most orbits. These results are confirmed in Table 7, which shows the number of orbits for which each of the statistical approaches was superior to the others.

3.4. Statistical retrievals using different fractions of deterministic AOT retrievals for training

Experiments described in Section 3.3 were performed using 10% of the deterministic AOT retrievals for training. We also repeated the exercise using 2%, 5%, 20%, and 50% of available deterministic retrievals as labeled data for training. This covers a wide range of retrieval speed tradeoffs. Since statistical retrievals are several orders of

Table 3
Hidden layer optimization for GANN

No. of hidden nodes	PCA retained variance					
	95% (31 attributes)			99% (54 attributes)		
	5	15	30	5	15	30
R^2 of cycle 1	0.825	0.839	0.825	0.839	0.842	0.840
R^2 of cycle 2	0.617	0.648	0.617	0.640	0.647	0.643
R^2 of cycle 3	0.353	0.359	0.353	0.357	0.372	0.375
R^2 of cycle 4	0.429	0.446	0.429	0.440	0.424	0.411

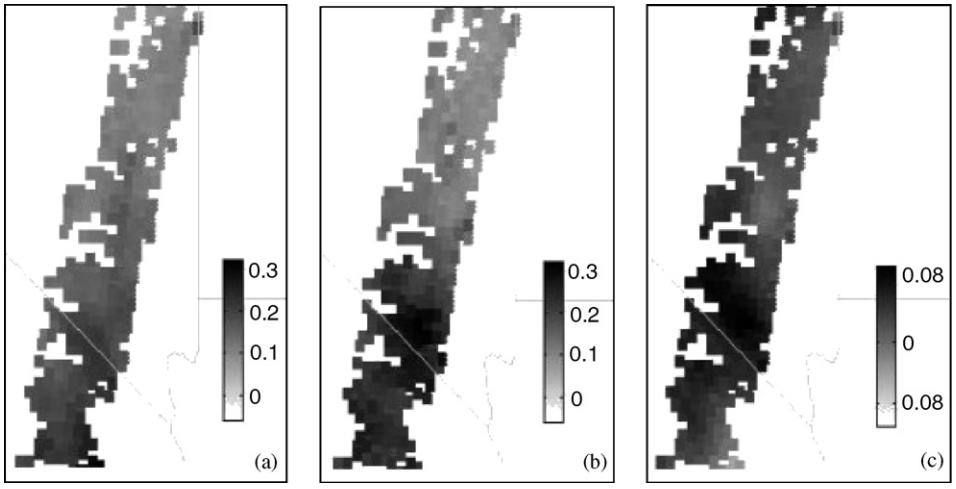

Fig. 2. (a) MISR AOT intensities retrieved over a portion of Orbit 7 (latitude$\in$[36.0, 41.7], longitude $\in$[−117.9, −112.1]) on July 10, 2002. (b) GANN AOT intensities retrieved over a portion of Orbit 7 (latitude$\in$[36.0, 41.7], longitude$\in$[−117.9, −112.1]) on July 10, 2002. (c) Errors of GANN AOT retrievals over a portion of Orbit 7 (latitude$\in$[36.0, 41.7], longitude $\in$[−117.9, −112.1]) on July 10, 2002.

Table 4
Accuracy of IDSI using different p values

p	0.5	1	1.5	2	2.5	3
R^2 of cycle 1	0.685	0.792	0.828	0.839	0.837	0.834
R^2 of cycle 2	0.395	0.504	0.561	0.648	0.607	0.586
R^2 of cycle 3	0.277	0.313	0.371	0.359	0.482	0.481
R^2 of cycle 4	0.318	0.398	0.433	0.446	0.434	0.426

Table 5
Comparison of overall accuracies achieved by different models

Model name	GANN	IDSI	RSNN	ECM	$WAM_{GANN+IDSI}$
R^2 of cycle 1	0.839	0.837	0.841	0.864	0.864
R^2 of cycle 2	0.648	0.582	0.465	0.700	0.728
R^2 of cycle 3	0.359	0.684	0.638	0.575	0.745
R^2 of cycle 4	0.446	0.439	0.465	0.472	0.661

magnitude faster than the deterministic one, using the deterministic method for 2% of locations and statistical algorithms over the remaining 98% results in a speed-up of almost 50 times.

By increasing the fraction of training data from 2% to 10%, retrieval accuracy was increased nearly linearly. Accuracy leveled off with a further increase in training set size (Fig. 3). We view this as computational support for the

Table 6
Paired model comparison

Comparison pair	Cycle 1	Cycle 2	Cycle 3	Cycle 4
WAM > GANN	27	25	26	14
WAM > IDSI	21	18	25	13
WAM > RSNN	28	29	24	14
WAM > ECM	30	27	18	13
ECM > GANN	19	12	20	13
ECM > IDSI	10	9	20	12
ECM > RSNN	22	22	22	13
IDSI > GANN	28	21	21	12
IDSI > RSNN	24	25	17	12
RSNN > GANN	16	7	12	4

Results show the number of orbits in which the first model had higher accuracy than the second model. Cycle 1 had 33, Cycle 2 had 30, Cycle 3 had 26, and Cycle 4 had 16 orbits.

Table 7
Number of orbits won by each model

Model name	GANN	IDSI	RSNN	ECM	$WAM_{GANN+IDSI}$
Cycle 1	3	9	2	3	16
Cycle 2	3	9	1	3	14
Cycle 3	0	0	4	7	15
Cycle 4	0	3	1	4	8

geoscientist's decisions to provide deterministic retrievals at 17.6 km × 17.6 km spatial resolution rather than at a higher resolution. Fig. 3 shows that over a large range of training data densities, statistical models obtained by weighted averaging of global and local predictors were superior in accuracy to global or local models by themselves. We also observed that ECM models followed GANN models: they performed better than both GANN and IDSI when GANN accuracies are large, and performed worse than IDSI when GANN accuracies are small. RSNN performance was comparable with IDSI when the training set size was large.

3.5. Quality comparison of statistical vs. deterministic retrievals

R^2 values reported in Tables 3–5 are not sufficient to conclude the extent to which statistical retrievals can be used to complement deterministic AOT retrieval algorithms in practice. In particular, in Table 5, smaller R^2 scores in cycle 4 are due to a very small standard deviation in the deterministic AOT retrievals in this period (Table 2) such that minimal prediction errors of statistical retrievals resulted in low R^2 scores. More insight is possible by comparing errors introduced by statistical AOT retrieval algorithms to those of deterministic algorithms.

MISR deterministic AOT retrieval algorithms have been undergoing extensive validation by comparison with ground-based AOT observations from Aerosol Robotic Network (AERONET) sites. A recent study (Abdou et al., 2005) reported that the standard deviation of MISR AOT retrieval error over land is approximately $0.05 \pm 0.2\tau$, where τ denotes actual AOT value. This indicates that the magnitude of AOT retrieval error increases with AOT. Using this, we estimate MSE of MISR deterministic AOT retrievals, $\mathrm{MSE}_{\mathrm{det}}$, as

$$\mathrm{MSE}_{\mathrm{det}} = \frac{1}{K}\sum_{i=1}^{K}(0.05 + 0.2 \times \mathrm{AOT}_i)^2. \quad (5)$$

Our regression models are constructed based on MISR AOT retrievals. Therefore, their prediction value y_{stat} can be represented as

$$y_{\mathrm{stat}} = y_{\mathrm{det}} + \varepsilon_{\mathrm{stat}} = y_{\mathrm{AERONET}} + \varepsilon_{\mathrm{det}} + \varepsilon_{\mathrm{stat}}. \quad (6)$$

Assuming independence between deterministic and statistical prediction errors, $\varepsilon_{\mathrm{det}}$ and $\varepsilon_{\mathrm{stat}}$, MSE of the statistical retrieval, MSE_{stat} can be estimated as

$$\begin{aligned}\mathrm{MSE}_{\mathrm{stat}} &= E[(\varepsilon_{\mathrm{det}} + \varepsilon_{\mathrm{stat}})^2] = E[\varepsilon_{\mathrm{det}}^2] + E[\varepsilon_{\mathrm{stat}}^2] \\ &= \mathrm{MSE}_{\mathrm{det}} + \mathrm{MSE}_{\Delta}, \end{aligned} \quad (7)$$

where MSE_{Δ} is MSE of our (statistical) prediction.

Based on formulas (5) and (7), we computed $\mathrm{MSE}_{\mathrm{stat}}$ for WAM and $\mathrm{MSE}_{\mathrm{det}}$ for deterministic AOT retrievals. Square root values of MSE (RMSE), commonly used in

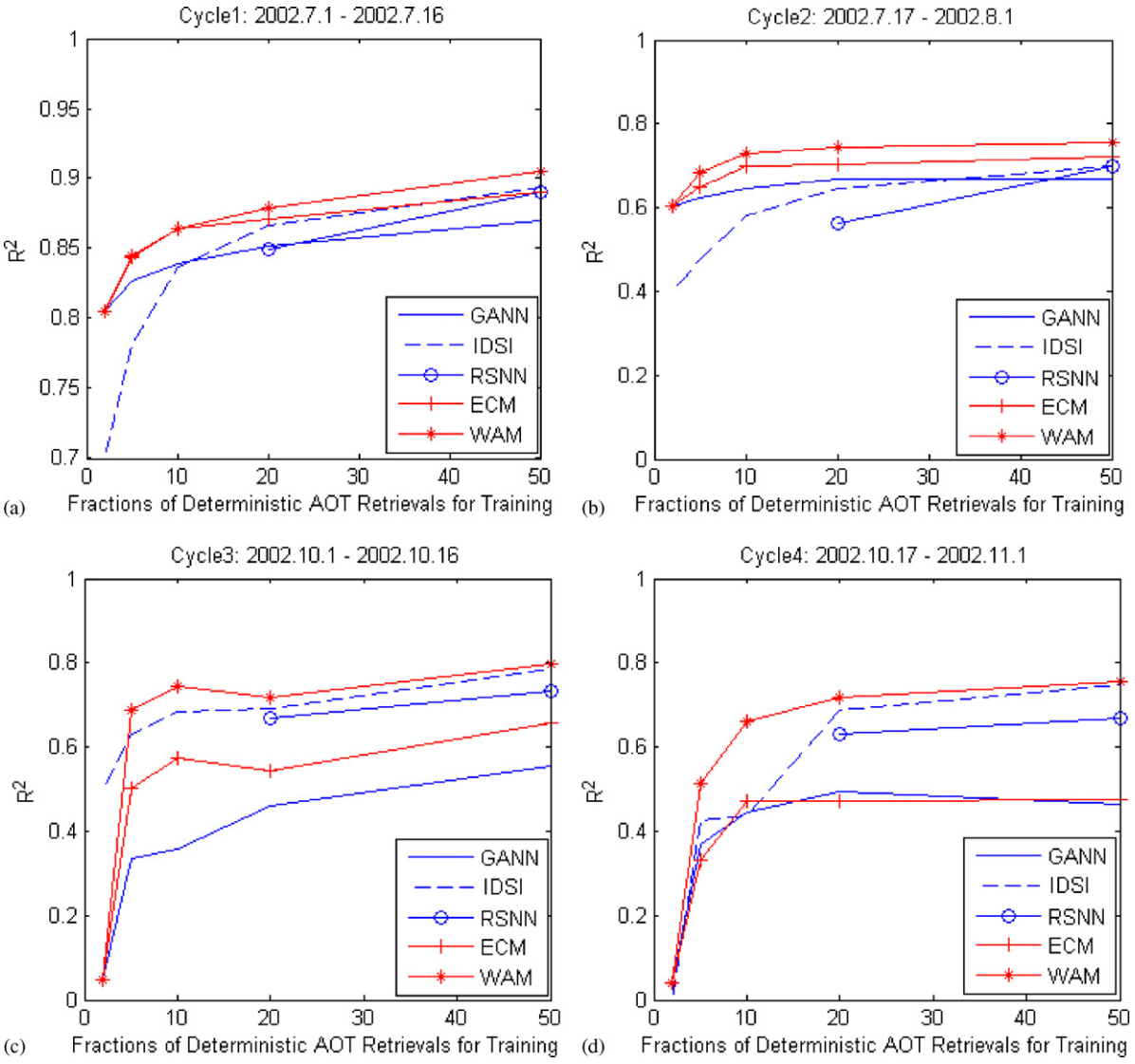

Fig. 3. Overall accuracy of statistical retrieval models GANN, IDSI, RSNN, ECM, and WAM when using 2–50% of available deterministically retrieved AOT for training of statistical models.

Table 8
Comparison of $RMSE_{stat}$ for WAM versus $RMSE_{det}$ of deterministic AOT retrievals over the same data

Type of algorithm	Statistical ($WAM_{GANN+IDSI}$)					Deterministic
	Fraction of training data					
	2%	5%	10%	20%	50%	
RMSE of cycle 1	0.149	0.143	0.139	0.137	0.131	0.113
RMSE of cycle 2	0.118	0.117	0.114	0.112	0.112	0.092
RMSE of cycle 3	0.078	0.074	0.072	0.073	0.071	0.068
RMSE of cycle 4	0.078	0.075	0.073	0.072	0.071	0.067

Results are shown for fractions of deterministic retrievals equal to 2%, 5%, 10%, 20%, and 50%.

the geosciences, are shown in Table 8. Statistical retrievals can be quite comparable in quality to deterministic ones, where the difference in quality varies with season and model complexity. For example, using deterministic retrieval over only 2% of locations, the $RMSE_{stat}$ was about 1.3 times larger than $RMSE_{det}$ in the first two cycles (summer period) and about 1.15 times larger than $RMSE_{det}$ in the last two cycles (autumn). By increasing the fraction to 20%, the difference in quality of decreased to 1.2 and 1.05 times, respectively.

4. Conclusion

In this study, we explored whether data-driven statistical retrievals can serve as an efficient and practically useful complement to traditional physical model-based deterministic retrieval methods. For statistical retrievals, we considered (1) global neural networks trained using a fraction of deterministic retrievals sampled from the entire domain, (2) local neural networks, and (3) local spatial interpolation models developed using data from limited regions (single orbits in our experiments). To address spatial heterogeneity of AOT data distribution, we proposed using local statistical models for correction of spatially correlated errors from global regressions. We also developed a weighted average-based ensemble model, which takes advantage of large global data sets but also exploits more specific spatial properties at local sites.

We evaluated this methodology in the context of AOT retrievals for four 16-day cycles of MISR data over the entire continental United States. In our experiments, both methods for integration of global and local statistical components were clearly superior to global and local statistical retrievals alone. The most accurate results were obtained through a weighted average optimization of global and local components. The benefits were particularly evident when a larger fraction (20%) of deterministic AOT retrievals were used for training, but relying on smaller fractions (2%) of deterministic retrievals also resulted in quite accurate results. This suggests that statistical AOT retrievals can serve as a practically useful complement to traditional deterministic retrieval methods in providing higher resolution retrievals with reduced computational efforts.

Acknowledgments

We thank Ralph A. Kahn and John Martonchik, MISR Science Team members at the Jet Propulsion Laboratory, for their generous help with understanding of the domain. We also thank the Atmospheric Sciences Data Center at NASA Langley Research Center for their support in collection of MISR data, and Qifang Xu and Yong Li at Temple University for data management and preprocessing help.

References

Abdou, W.A., Diner, D.J., Martonchik, J.V., Bruegge, C.J., Kahn, R.A., Gaitley, B.J., Crean, K.A., 2005. Comparison of coincident Multiangle Imaging Spectroradiometer and Moderate Resolution Imaging Spectroradiometer aerosol optical depths over land and ocean scenes containing aerosol robotic network sites. Journal of Geophysical Research 110, 11967–11976.

Berdnik, V.V., Loiko, V.A., 2006. Particle sizing by multiangle light-scattering data using the high-order neural networks. Journal of Quantitative Spectroscopy and Radiative Transfer 100, 55–63.

Bothwell, G.W., Hansen, E.G., Vargo, R.E., Miller, K.C., 2002. The Multi-angle Imaging SpectroRadiometer science data system, its products, tools, and performance. IEEE Transactions on Geoscience and Remote Sensing 40 (7), 1467–1476.

Bull, M., Matthews, J., Moroney, C., Smyth, M., 2005. Multi-angle imaging SpectroRadiometer data product specifications. Technical Report, Jet Propulsion Laboratory at California Institute of Technology.

Diner, D.J., 1999. Multi-angle imaging SpectroRadiometer (MISR) experiment overview. Technical Report, Jet Propulsion Laboratory at California Institute of Technology.

Diner, D.J., Davies, R., 2003. Multi-angle imaging of the earth: present and future. Technical Report, Jet Propulsion Laboratory at California Institute of Technology.

EOS Data Gateway, 2005. http://delenn.gsfc.nasa.gov/~imswww/pub/imswelcome/.

Faure, T., Isaka, H., Guillemet, B., 2002. Neural network retrieval of cloud parameters from high-resolution multispectral radiometric data: a feasibility study. Remote Sensing of Environment 80 (2), 285–296.

Han, B., Vucetic, S., Braverman, A., Obradovic, Z., 2005. Integration of deterministic and statistical algorithms for aerosol retrieval. In: Proceedings of the Ninth International Conference on Engineering Applications of Neural Networks, Lille, France, pp. 85–92.

Herring, D.D., King, M.D., 2000. Space-based observations of the earth. In: Murdin, K. (Ed.), Encyclopedia of Astronomy and Astrophysics. Institute of Physics Publishing, pp. 2959–2962.

Kahn, R.A., Diner, D.J., Martonchik, J.V., West, R.A., 1997. Multiangle remote sensing of aerosols over ocean. Technical Report, Jet Propulsion Laboratory at California Institute of Technology.

Martonchik, J.V., Diner, D.J., Kahn, R.A., Ackerman, T.P., Verstraete, M.M., Pinty, B., Gordon, H., 1998. Techniques for the retrieval of aerosol properties over land and ocean using multiangle imaging. IEEE Transactions on Geoscience and Remote Sensing 36 (4), 1212–1227.

Martonchik, J.V., Diner, D.J., Crean, K.A., Bull, M.A., 2002. Regional aerosol retrieval results From MISR. IEEE Transactions on Geoscience and Remote Sensing 40 (7), 1520–1531.

Mcguckin, B.T., Haner, D.A., Menzies, R.T., 1995. Multi-angle imaging SpectroRadiometer (MISR): optical characterization of the spectralon calibration panels. Technical Report, Jet Propulsion Laboratory at California Institute of Technology.

Ramanathan, V., Crutzen, P.J., Kiehl, J.T., Rosenfeld, D., 2002. Atmosphere–'aerosols, climate, and the hydrological cycle'. Science 294 (5549), 2119–2124.

Vucetic, S., Fiez, T., Obradovic, Z., 2000. Examination of the influence of data aggregation and sampling density on spatial estimation. Water Resources Research 36 (12), 3721–3731.

Zhao, M., Heinsch, F.A., Nemani, R.R., Running, S.W., 2005. Improvements of the MODIS terrestrial gross and net primary production global dataset. Remote Sensing of Environment 95 (2), 164–176.