Machine Learning Applications for Earth Observation

David J. Lary, Gebreab K. Zewdie, Xun Liu, Daji Wu, Estelle Levetin, Rebecca J. Allee, Nabin Malakar, Annette Walker, Hamse Mussa, Antonio Mannino, and Dirk Aurin

Abstract Machine learning has found many applications in remote sensing. These applications range from retrieval algorithms to bias correction, from code acceleration to detection of disease in crops, from classification of pelagic habitats to rock type classification. As a broad subfield of artificial intelligence, machine learning is concerned with algorithms and techniques that allow computers to "learn" by example. The major focus of machine learning is to extract information from data automatically by computational and statistical methods. Over the last decade there has been considerable progress in developing a machine learning methodology for a variety of Earth Science applications involving trace gases, retrievals, aerosol products, land surface products, vegetation indices, and most recently, ocean applications. In this chapter, we will review some examples of how machine learning has already been useful for remote sensing and some likely future applications.

Introduction

Beyond remote sensing, machine learning has already proved immensely useful in a wide variety of applications in science, business, health care, and engineering. Machine learning allows us to *learn by example*, and to *give our data a voice*. It is particularly useful for those applications for which we do *not* have a complete theory, yet which are of significance. Machine learning is an automated implementation of the scientific method (Domingos, 2015), following the same process of generating, testing, and discarding or refining hypotheses. While a scientist or engineer may spend his entire career coming up with and testing a few hundred hypotheses, a machine-learning system can do the same in a fraction of a second. Machine learning provides an objective set of tools for automating discovery. It is

D.J. Lary (✉) • G.K. Zewdie • X. Liu • D. Wu • E. Levetin • R.J. Allee • N. Malakar
A. Walker • H. Mussa • A. Mannino • D. Aurin
Hanson Center for Space Sciences, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080, USA
e-mail: david.lary@utdallas.edu; https://davidlary.info

P.-P. Mathieu, C. Aubrecht (eds.), *Earth Observation Open Science and Innovation*,
ISSI Scientific Report Series 15, https://doi.org/10.1007/978-3-319-65633-5_8

therefore not surprising that machine learning is currently revolutionizing many areas of science, technology, business, and medicine (Lary et al., 2016).

Machine learning is now being routinely used to work with large volumes of data in a variety of formats such as image, video, sensor, health records, etc. Machine learning can be used in understanding this data and creating predictive and classification tools. When machine learning is used for regression, empirical models are built to predict continuous data, facilitating the prediction of future data points, e.g. algorithmic trading and electricity load forecasting. When machine learning is used for classification, empirical models are built to classify the data into different categories, aiding in the more accurately analysis and visualization of the data. Applications of classification include facial recognition, credit scoring, and cancer detection. When machine learning is used for clustering, or unsupervised classification, it aids in finding the natural groupings and patterns in data. Applications of clustering include medical imaging, object recognition, and pattern mining. Object recognition is a process for identifying a specific object in a digital image or video. Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques. These technologies are being used for applications such as driver-less cars, automated skin cancer detection, etc.

There are now a variety of open source tools that can greatly facilitate the use of machine learning, such as scikit-learn,[1] TensorFlow,[2] Caffe,[3] and Spark Mlib.[4] Common programming environments used for machine learning include R,[5] Python,[6] and Matlab.[7] All of the applications shown in this chapter used matlab.

In this paper we will give an overview of several remote sensing applications of machine learning made over the last two decades and then take a look ahead to some likely future applications.

What Is Machine Learning?

Machine learning is an automated approach to building empirical models from the data *alone*. A key advantage of this is that we make *no* a priori assumptions about the data, its functional form, or probability distributions. It is an empirical approach, so we do not need to provide a theoretical model. However, it also means that for machine learning to provide the best performance we do need a

[1] http://scikit-learn.org/stable/.

[2] https://www.tensorflow.org.

[3] http://caffe.berkeleyvision.org.

[4] http://spark.apache.org/mllib/.

[5] https://cran.r-project.org.

[6] https://www.python.org.

[7] https://www.mathworks.com/solutions/machine-learning.html.

comprehensive representative set of examples, that spans as much of the parameter space as possible. This comprehensive set of examples is referred to as the *training data*.

So, for a successful application of machine learning we have *two* key ingredients, both of which are essential, a machine learning algorithm, and a comprehensive training data set. Then, once the training has been performed, we should test its efficacy using an independent validation data set to see how well it performs when presented with data that the algorithm has *not* previously seen, i.e. test its *generalization*. This can be, for example, a randomly selected subset of the training data that was held back and then utilized for independent validation.

It should be noted that with a given machine learning algorithm, the performance can go from poor to outstanding with the provision of a progressively more complete training data set. Machine learning really is learning by example, so it is critical to provide as complete a training data set as possible. At times, this can be a labor-intensive endeavor.

When using machine learning we are typically performing one of three tasks:

1. Multivariate non-linear non-parametric regression.
2. Supervised classification.
3. Unsupervised classification.

Each of these tasks can be achieved by a variety of different algorithms. Some of the commonly used algorithms include Neural Networks (McCulloch and Pitts, 1943; Haykin, 2001, 2007, 1994, 1999; Demuth et al., 2014; Bishop, 1995), Support Vector Machines (Vapnik, 1982, 1995, 2000, 2006; Cortes and Vapnik, 1995), Decision Trees (Safavian and Landgrebe, 1991), and Random Forests (Ho, 1998; Breiman, 1984, 2001).

Let us now turn our attention to some examples.

Some Existing Machine Learning Applications

We will start by looking at several examples of bias correction. Bias identification and correction is of particular importance for every single remote sensing instrument. Bias correction can also prove to be a particularly challenging issue, one which involves multiple factors.

Machine Learning for Bias Correction and Cross Calibration

The ubiquitous issue of inter-instrument biases is an obvious example of where we do *not* have a complete theoretical understanding, and so machine learning can be of particular use.

In many areas of remote sensing we have multiple instruments simultaneously observing the earth on a variety of platforms. Many of these sensors may be providing data on the same parameters, such as the surface vegetation, the composition of the atmosphere or ocean. A ubiquitous issue faced is inter-instrument bias between the contemporaneously observing instruments. This inter-instrument bias can be due to a variety of known reasons that may include different instruments, different observing geometry and orbits, etc., as well as some causes that we do not know.

This is an important issue, as we routinely need to provide data fusion of multiple datasets. Datasets which are inevitably biased relative to each other, sometimes even after the mandatory calibration/validation process. When we are seeking to construct a long-term record spanning many decades this inevitably will often involve a large number of instruments, a matter very relevant for climate variables. In addition, Data Assimilation has become an important part of effectively utilizing remotely sensed data. However, data assimilation is a *Best Linear Unbiased Estimator* (BLUE), and fusing biased data can cause serious issues.

This data fusion typically involves large teams of scientists and engineers. On the one hand, the instrument teams have a keen sense of faithfully reporting the data, as it is, warts and all. They are naturally loath to empirically correct biases; they would like to theoretically understand the cause of the bias and data issues from first principles. However, as the Earth System is so complex, with many interacting processes, and often the instruments are also complex, this is not always possible. Residual data issues can, and usually do, remain. On the other hand, the modelers know that data bias exists, but are very reticent to make changes to data products that they did not collect, so we therefore have *a problem of closure*.

Biases are ubiquitous, not all of them can be explained theoretically. Yet, we typically need to fuse multiple datasets to construct long-term time series and/or improve global coverage. If the biases are not corrected before data fusion we introduce further problems, such as spurious trends, leading to the possibility of unsuitable policy decisions. When data assimilation is involved, any use of biased observations can lead to the sub-optimal use of the observations, non-physical structures in the analysis, biases in the assimilated fields, and extrapolation of biases due to multivariate background constraints. To compound matters further, the instruments whose data we would like to fuse are often not making coincident measurements in time or space. It is imperative to inter-compare observations in their appropriate context and be able to address the pernicious issue of inter-instrument bias. An issue where machine learning has proved to be most useful. Let us now take a look at some examples.

Vegetation Indices

Consistent, long-term vegetation data records are critical for analysis of the impact of global change on terrestrial ecosystems. Continuous observations of terrestrial ecosystems through time are necessary to document changes in magnitude or variability in an ecosystem. Satellite remote sensing has been the primary tool for

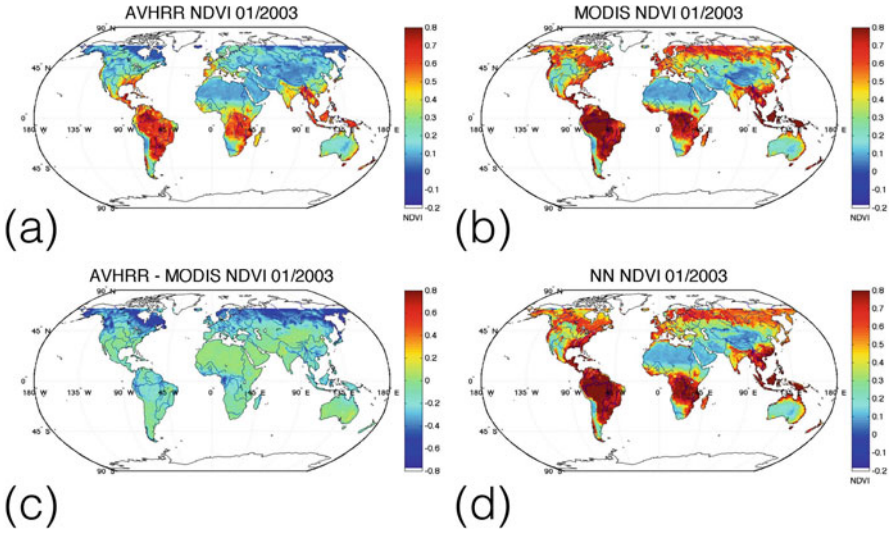

Fig. 1 Using the same color scale, panels (**a**) and (**b**) show the contemporaneous January 2003 NDVI averages for AVHRR and MODIS, respectively. The differences are particularly evident in the higher northern latitudes. Panel (**c**) shows the absolute difference in NDVI between AVHRR and MODIS. Panel (**d**) shows the *estimated* MODIS NDVI for January 2003 using AVHRR data and a neural network. The neural network has been extremely effective in removing the substantial bias in NDVI between AVHRR and MODIS

scientists to measure global trends in vegetation, as the measurements are both global and temporally frequent. To extend measurements through time, multiple sensors with different design and resolution must be used together in the same time series. This presents significant problems as sensor band placement, spectral response, processing, and atmospheric correction of the observations can vary significantly and impact the comparability of the measurements.

Even without differences in atmospheric correction, vegetation index values for the same target recorded under identical conditions will not be directly comparable because input reflectance values differ from sensor to sensor due to differences in sensor design and spectral response of the instrument. This is clearly visible in the example shown in Fig. 1. Using the same color scale, panels (a) and (b) show the contemporaneous January 2003 NDVI averages for AVHRR and MODIS, respectively. The differences are particularly evident in the higher northern latitudes. Panel (c) shows the difference in NDVI between AVHRR and MODIS, there are substantial biases present.

Brown et al. (2008) showed how machine learning, in particular, neural networks, can identify and remove differences in sensor design and variable atmospheric contamination from the AVHRR NDVI record in order to match the range and variance of MODIS NDVI without removing the desired signal representing the underlying vegetation dynamics. This is well illustrated by comparing Fig. 1 panels (b) and (d). Panel (b) shows the actual MODIS NDVI for January 2003. Panel

(d) shows the *estimated* MODIS NDVI for January 2003 using AVHRR data and a neural network, they are almost indistinguishable. The neural network has been extremely effective in removing the substantial bias in NDVI between AVHRR and MODIS.

Neural networks are "data transformers," where the objective is to associate the elements of one set of data to the elements in another. Relationships between the two datasets can be complex and the two datasets may have different statistical distributions. This transformation incorporates additional input data that may account for differences between the two datasets.

Brown et al. (2008) demonstrated the viability of neural networks as a tool to produce a long-term dataset based on AVHRR NDVI that has the data range and statistical distribution of MODIS NDVI. Previous work has shown that the relationship between AVHRR and MODIS NDVI is complex and nonlinear, thus this problem is well suited to neural networks if appropriate inputs can be found. The impact of atmospheric contamination, such as clouds, smoke, pollution, and other aerosols, variations in soil color and exposure through vegetation, and land cover type has a differential effect on AVHRR data as compared to MODIS data. Brown et al. (2008) used overlapping years of observations to train the neural networks.

Remote sensing datasets are the result of a complex interaction between the design of a sensor, the spectral response function, stability in orbit, the processing of the raw data, composting schemes, and post-processing corrections for various atmospheric effects including clouds and aerosols. The interaction between these various elements is often nonlinear and non-additive, where some elements increase the vegetation signal-to-noise ratio (compositing, for example) and others reduce it (clouds and volcanic aerosols). Thus, although many have used simulated data to explore the relationship between AVHRR and MODIS, these techniques are not directly useful in producing a sensor-independent vegetation dataset that can be used by data users in the near term.

There are substantial differences between the processed vegetation data from AVHRR and MODIS. In order to have a long data record that utilizes all available data back to 1981, we must find practical ways of incorporating the AVHRR data into a continuum of observations that include both MODIS and VIIRS. Brown et al. (2008) showed that the TOMS data record on clouds, ozone, and aerosols can be used to identify and remove sensor-specific atmospheric contaminants that differentially affect the NDVI from AVHRR over MODIS. Other sensor-related effects, particularly those of changing BRDF, viewing angle, illumination, and other effects that are not accounted for here, remain important sources of additional variability. Although this analysis has not produced a dataset with identical properties to MODIS, it has demonstrated that a neural net approach can remove most of the atmospheric-related aspects of the differences between the sensors, and match the mean, standard deviation, and range of the two sensors. A similar technique can be used for the VIIRS sensor.

Let us now look at some other examples related to the remote sensing of atmospheric composition.

Space-Based Measurements of HCl Relevant for Ozone Depletion

The magnitude of atmospheric ozone depletion is closely intertwined with the abundance of atmospheric halogens such as chlorine. The main reservoir for atmospheric chlorine compounds is HCl. The peak in stratospheric HCl was reached in the late 1990s. Between 1998 and 2004 the stratospheric loading of HCl was relatively constant, with some month to month fluctuation; this was followed by a

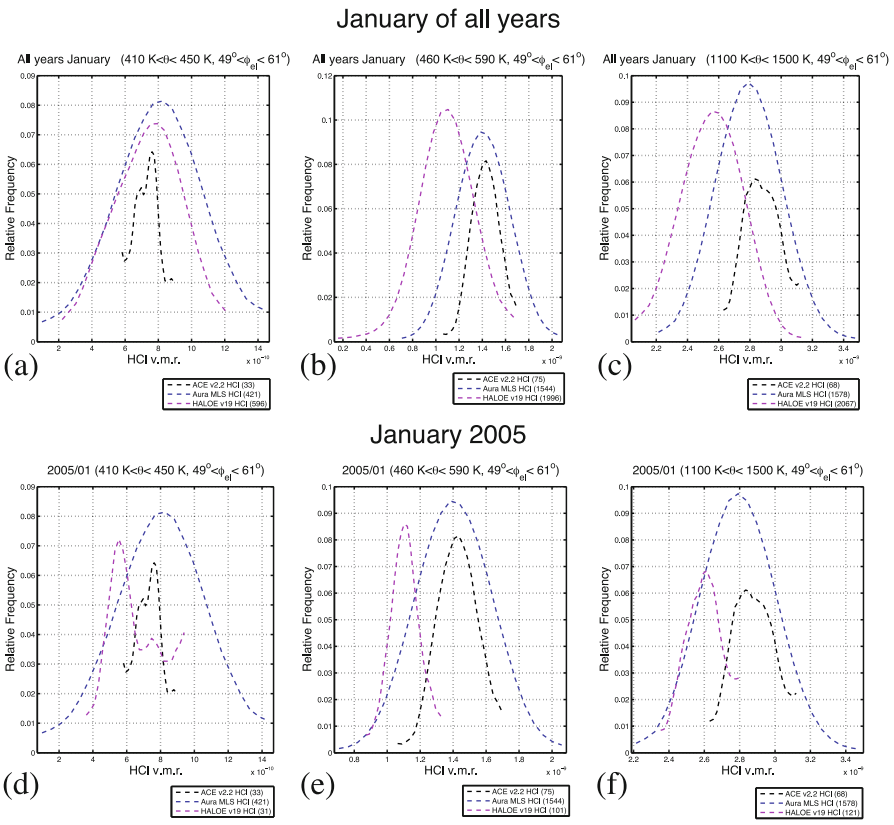

Fig. 2 Example HCl PDFs for the three instruments HALOE, ACE, and Aura MLS. In each case the PDFs are for all observations made by that instrument in a Lagrangian region for three isentropic levels centered on an equivalent latitude of 55°N during all the Januarys that the instrument observed. For ACE the plots include January 2004–2006, for HALOE the plots include 2004–2005, and for MLS the plots include 2005–2006. (**a**) Plot of a PDF for all observations in the range $410\,\mathrm{K} < \theta < 450\,\mathrm{K}$ ($70\,\mathrm{mbar} < P < 110\,\mathrm{mbar}$), $49° < \phi_{el} < 61°$. (**b**) Plot of a PDF for all observations in the range $460\,\mathrm{K} < \theta < 590\,\mathrm{K}$ ($30\,\mathrm{mbar} < P < 60\,\mathrm{mbar}$), $49° < \phi_{el} < 61°$. (**c**) Plot of a PDF for all observations in the range $1100\,\mathrm{K} < \theta < 1500\,\mathrm{K}$ ($2\,\mathrm{mbar} < P < 4\,\mathrm{mbar}$), $49° < \phi_{el} < 61°$. Panels (**d**)–(**f**) are analogous to panels (**a**)–(**c**) for the observations made only during January 2005. The number of observations used to form each PDF is shown in parentheses in the legend

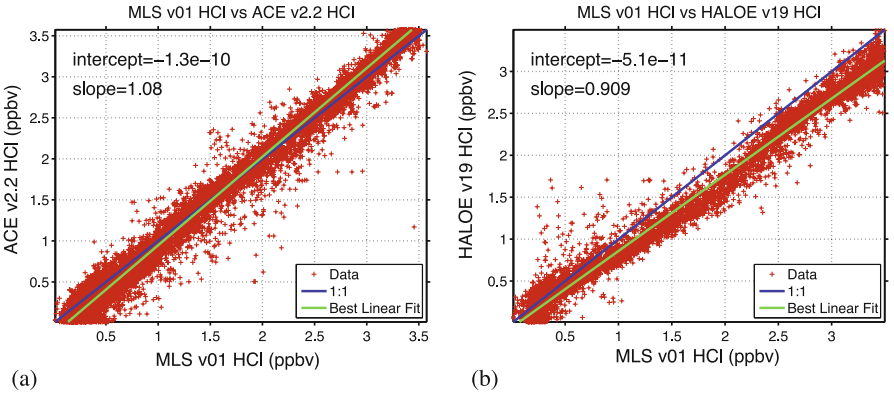

Fig. 3 Panels (**a**) and (**b**) show scatter plots of all contemporaneous observations of HCl made by HALOE, ACE, and Aura MLS. Each point plotted is the median value of a PDF of observations made for a Lagrangian region over the period of a month. It can be seen that to adequately using the slopes to describe the differences does not do justice to the differences. For example, panel (**a**) shows a much better agreement than panel (**b**), but the slopes themselves do not reflect this. Panel (**b**) shows an offset of order 9% near 2 ppbv, whereas panel (**a**) shows maybe 1 or 2% for values near 2 ppbv. The mean absolute value of the differences seems a good indicator of the fits

more pronounced decrease in HCl since 2004. As can be seen in Figs. 2 and 3, we can use probability distribution functions (PDFs) and scatter diagrams for validation and bias characterization of Aura Microwave Limb Sounder (MLS) HCl retrievals. Both these methods allow us to use large statistical samples and do not require correlative measurements to be collocated in space and time.

We can take the difference between the medians of the PDFs as a measure of the inter-instrument bias. This bias is really only significant if it is larger than the atmospheric variability in the Lagrangian region we are considering (i.e., the width of the PDF).

We compared the PDFs for all overlapping Lagrangian regions for a given month. However, we can use a single scatter diagram to compare all the overlaps globally for all the months observed by each pair of instruments (Fig. 3). Such a scatter diagram has the advantage of a *huge sample size*, it encompasses the entire period that a pair of instruments were making contemporaneous observations. The scatter diagram is intended as a big picture summary for all contemporaneous observations made globally. Figure 4 shows two scatter diagrams for all the contemporaneous observations of HCl made globally by two pairs of instruments. In Fig. 4a we compare ACE and Aura MLS which were making contemporaneous observations between September 2004 and the present. In Fig. 4b we compare HALOE and Aura MLS which were making contemporaneous observations between September 2004 and November 2005.

In the ideal case where we have perfect agreement between two instruments, the slope of the scatter diagram would be 1 and the intercept would be 0. In the case of ACE and Aura, we see there is a slope of 1.08, and for HALOE Aura there is a

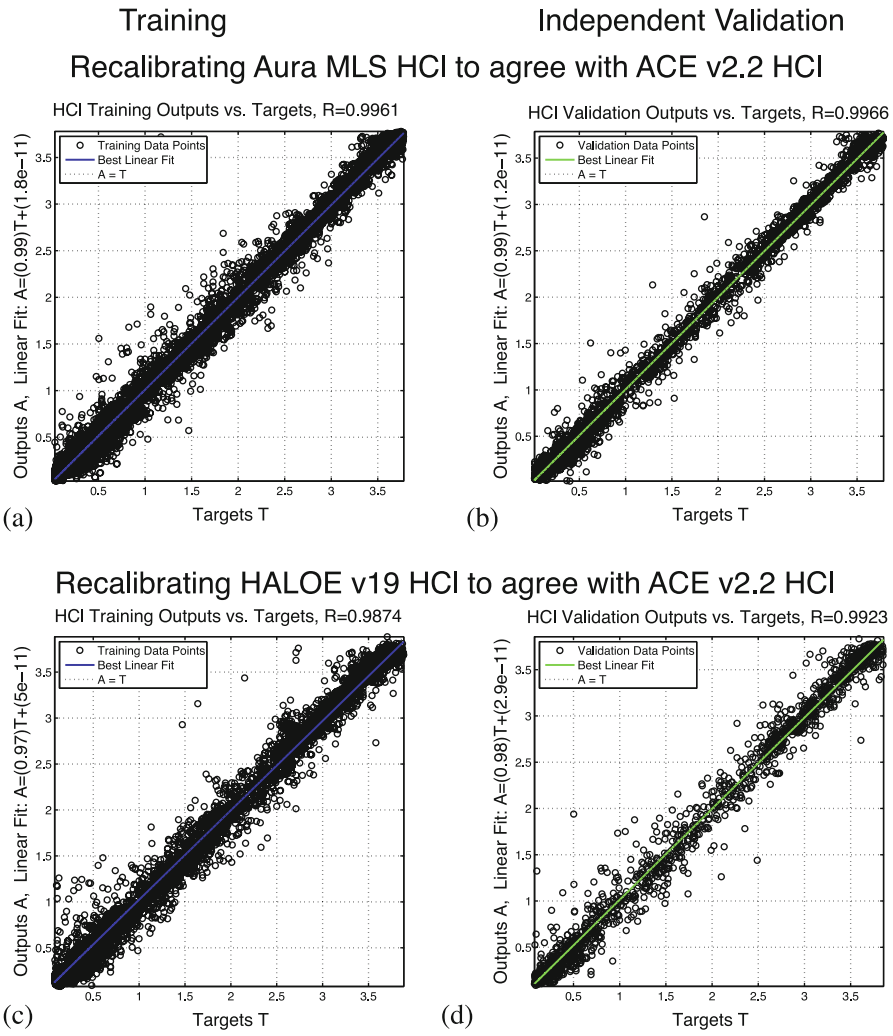

Fig. 4 (left) Result of training a neural network to learn the interinstrument biases. (right) An independent validation of this training using a randomly chosen, totally independent, data sample not used in training the neural network. In each case, the *x* axis shows the actual ACE v2.2 HCl (the target). (**a**) and (**b**) The *y* axis is the neural network estimate of ACE v2.2 HCl based on Aura MLS v01 HCl. Panel (**a**) is the result using the training data, and Fig. 5b is the result of the independent validation. Panels (**c**) and (**d**) The *y* axis is the neural network estimate of ACE v2.2 HCl based on HALOE v19 HCl. Figure 5c is the result using the training data, and panel (**d**) is the result of the independent validation. In each case, this is a global training for all contemporaneous observations between each pair of instruments. The training points are the median values of a PDF of observations made during a given month for a given equivalent PV latitude—potential temperature bin. The width of the cloud of points in each of these scatter diagrams is a good measure of the uncertainty associated with the neural network fit

slope of 0.91 (Fig. 3). It can be seen that solely using the slopes does not do justice to the differences. For example, Fig. 3a shows a much better agreement than Fig. 3b, but the slopes themselves do not reflect this. Figure 3b shows an offset of order 9% near 2 ppbv, whereas Fig. 3a shows maybe 1 or 2% for values near 2 ppbv. The mean absolute value of the differences seems a good indicator of the fits. We also note that in the case of Aura MLS and HALOE, the scatter diagrams do not have a constant slope over the entire range of HCl values, several "wiggles" are present. This means that the inter-instrument biases are *spatially and temporally dependent*. Neural networks are multi-variate, non-parametric, "learning" algorithms that are ideally suited to learning, and correcting for, such inter-instrument biases.

We have used a neural network with three inputs and one output. The inputs are equivalent PV latitude, potential temperature, and HCl from instrument A. The output is HCl from instrument B. Potential temperature and equivalent latitude are used because they are good markers of the large-scale flow pattern. When we do the training we randomly split our training data set into three portions of 80%, 10%, and 10%. The 80% is used to train the neural network weights. This training is iterative and on each iteration we evaluate the current RMS error of the neural network. The RMS error is calculated by using the second 10% of the data that was not used in the training. We use the RMS error and the way it changes with training iteration (epoch) to determine the convergence of our training. When the training is complete, we use the final 10% as a validation data set. This 10% of the data was randomly chosen and not used in either the training or RMS evaluation. We only use the neural network if the validation scatter diagram, which plots the actual data from validation portion against the neural network estimate, yields a straight line graph with a slope of 1. This is a stringent and independent validation. The validation is global as the data was randomly selected over all temporal and spatial data points available. Several training strategies were examined, the one described included the most species over the longest time period. The neural network algorithm used was a feed-forward back-propagation network with 20 hidden nodes. The training was done by the Levenberg-Marquardt back-propagation algorithm.

Figure 4 shows the results of such a neural network training to learn inter-instrument biases between ACE v2.2, Aura MLS v1 and HALOE v19 HCl. Panels (b) and (d) show an independent validation of the training using a randomly chosen, totally independent, data sample not used in training the neural network. In each case the x axis shows the actual ACE v2.2 HCl (the target). In panels (a) and (b) the y axis is the neural network estimate of ACE v2.2 HCl based on Aura MLS v01 HCl. Panel (a) shows the results using the training data, panel (b) shows the results of the independent validation. In panels (c) and (d) the y axis is the neural network estimate of ACE v2.2 HCl based on HALOE v19 HCl. Panel (c) shows the results using the training data, and panel (d) shows the results of the independent validation. The mapping has removed the bias between the measurements and has also straightened out the "wiggles" seen in Fig. 3.

The bias between the Halogen Occultation Experiment (HALOE) and Aura MLS is greatest above the 525 K (21 km) isentropic surface. The global average mean bias between Aura and the Atmospheric Chemistry Experiment (ACE) for January 2005

was 2% and between Aura MLS and HALOE was 14%. The widths of the PDFs are a measure of the spatial variability and measurement precision. The Aura MLS HCl PDFs are consistently wider than those for ACE and HALOE, this reflects the retrieval uncertainties. The median observation uncertainty for Aura MLS v1.51 HCl is 12%, and the median ACE v2.2 uncertainty is 8%. We also connect Aura MLS HCl with the heritage of HALOE HCl by using neural networks to learn the inter-instrument biases and provide a seamless HCl record from the launch of the Upper Atmosphere Research Satellite (UARS) in 1991 to the present (Lary and Aulov, 2008).

HCl and Cl_y Time Series

Knowledge of the distribution of inorganic chlorine Cl_y in the stratosphere is needed to attribute changes in stratospheric ozone to changes in halogens, and to assess the realism of chemistry-climate models. However, there are limited direct observations of Cl_y. Simultaneous measurements of the major inorganic chlorine species are rare. In the upper stratosphere, Cl_y can be inferred from HCl alone.

Now that we have completely characterized the inter-instrument biases and been able to correct for them we can connect Aura MLS HCl observations to the heritage of HALOE (Lary et al., 2007). This allows us to produce an HCl time series from the launch of UARS in 1991 up to the present. Figure 5 shows HCl time series for six different locations with HCl observations from HALOE, ATMOS, ACE, MkIV and Aura MLS.

The HCl re-calibrations have been used (Fig. 6) to form a long Cl_y time series and associated uncertainty estimate (typically 0.4 ppbv at 800 K). The uncertainty in the Cl_y estimate is primarily due to the discrepancy between the different observations of HCl, i.e., the HALOE, Aura MLS, and ACE inter-instrument biases.

A consistent time series of stratospheric Cl_y from 1991 (Fig. 6) has been formed using available space-borne observations (Lary et al., 2007). Here we used neural networks to inter-calibrate HCl measurements from different instruments, and to estimate Cl_y from observations of HCl and CH_4. These estimates of Cl_y peaked in the late 1990s and have begun to decline as expected from tropospheric measurements of source gases and troposphere to stratosphere transport times. Furthermore, the estimates of Cl_y are consistent with calculations based on tracer fractional releases and age of air. The Cl_y time-series formed here is an important benchmark for models being used to simulate the recovery of the ozone hole. Although there is uncertainty in the estimates of Cl_y, primarily due to biases in HCl measurements, this uncertainty is small compared with the range of model predictions shown in the 2006 WMO report (Lary et al., 2007). This work was viewed as ground breaking and received three awards as a JCET science highlight, a NASA Aura Mission Science highlight, and as a NASA GSFC Atmospheric Chemistry and Dynamics Branch selected publication.

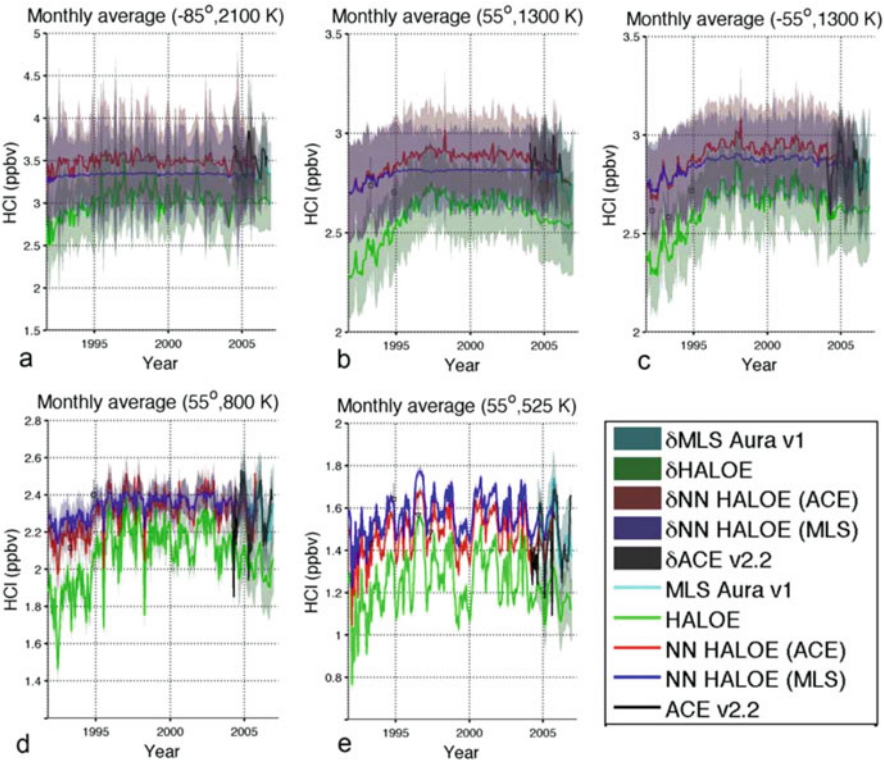

Fig. 5 Selected HCl time series from the launch of UARS to 2007 with HCl observations from HALOE, ATMOS, ACE, and Aura. (**a**) For 2100 K (50 km) at 85°S, (**b**) and (**c**) for 1300 K (41 km) at 55°N and 55°S, and (**d**) and (**e**) for 55°N at 800 K (30 km) and 525 K (21 km). In each case the green line and shading is for the original HALOE v19 data, and the red line and shading is for HALOE data remapped with a neural network to agree with Aura MLS v1 HCl. The black line is the ACE v2.2 data with the grey shading representing the associated uncertainty. The d in the legend which labels the shading refers to the total uncertainty (observational, representativeness, and where relevant, neural network adjustment). The remapping of HALOE generally brings the HALOE data into better agreement with the ATMOS (squares) data

Bias Correction of MODIS Aerosol Optical Depth

Aerosol and cloud radiative effects remain some of the largest uncertainties in our understanding of climate change. Over the past two decades, observations and retrievals of aerosol characteristics have been conducted from space-based sensors, from airborne instruments, and from ground-based samplers and radiometers. Much effort has been directed at these data sets to collocate observations and retrievals and to compare results. Ideally, when two instruments measure the same aerosol characteristic at the same time, the results should agree within well-understood measurement uncertainties. When inter-instrument biases exist, we would like to explain them theoretically from first principles. One example of this task is the

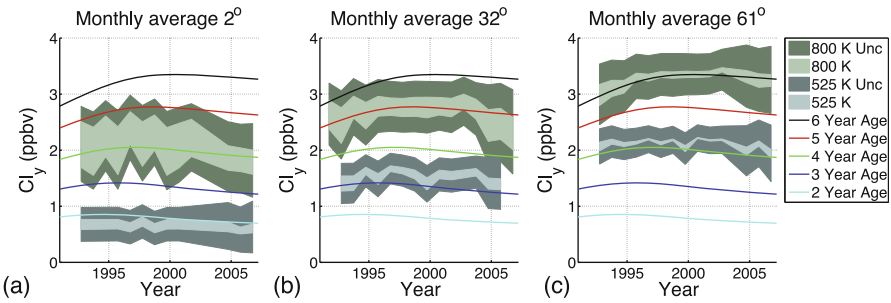

Fig. 6 (**a–c**) October Cl_y time-series for the 525 K isentropic surface (≈20 km) and the 800 K isentropic surface (≈30 km). In each case the dark shaded range represents the total uncertainty in our estimate of Cl_y. This total uncertainty includes the observational uncertainty, the representativeness uncertainty (the variability over the analysis grid cell), the inter-instrument bias in HCl, the uncertainty associated with the neural network inter-instrument correction, and the uncertainty associated with the neural network inference of Cl_y from HCl and CH_4. The inner light shading depicts the uncertainty on Cly due to the inter-instrument bias in HCl alone. The upper limit of the light shaded range corresponds to the estimate of Cl_y based on all the HCl observations calibrated by a neural network to agree with ACE v2.2 HCl. The lower limit of the light shaded range corresponds to the estimate of Cl_y based on all the HCl observations calibrated to agree with HALOE v19 HCl. Overlaid are lines showing the Cl_y based on age of air calculations. To minimize variations due to differing data, coverage months with less than 100 observations of HCl in the equivalent latitude bin were left out of the time-series

comparison between the aerosol optical depth (AOD) retrieved by the MODerate resolution Imaging Spectroradiometer (MODIS) and the AOD measured by the Aerosol Robotic Network (AERONET). While progress has been made in understanding the biases between these two data sets, we still have an imperfect understanding of the root causes.

The MODIS instruments are aboard both the Aqua and Terra satellites, launched on May 4, 2002 and December 18, 1999, respectively. The MODIS instruments collect data over the entire globe in 2 days. The AOD is retrieved using dark target methods in bands at 550, 670, 870, 1240, 1630, and 2130 nm, over ocean, and at 470, 550, and 670 nm over land. Other wavelengths are also used in the retrieval, for instance, short-wave infrared wavelengths for the land algorithm. Previous MODIS aerosol validation studies have compared the Aqua and Terra MODIS-retrieved AOD with the ground-based AERONET observations. AERONET is a global system of ground-based sun and sky scanning sun photometers that measure AOD in various channels, depending on individual instrument, but usually include measurements at 340, 380, 440, 500, 675, 870, and 1020 nm. Measurements are taken every 15 min during daylight hours. AERONET Level 2 quality assured AOD observations are accurate to within 0.01 for wavelengths of 440 nm and higher.

Previous studies concluded that MODIS AOD agreed with AERONET observations to within MODIS expected uncertainties, on a global basis. AERONET is only available for land locations, although some sites are in coastal regions. However, the correlation for the MODIS ocean algorithm was much better than

the agreement for the MODIS land algorithm, in the Collection 4 data set. Revision and implementation of a new land algorithm and reprocessing of the data resulted in much improvement to the retrieved MODIS AOD over land. Even so, there remains a small overprediction of the AOD for low values and underprediction at high AOD values.

Data Description

Lary et al. (2009) used the global 10 km MODIS Collection 5 AOD product, over land and ocean, and all the available AERONET version 2.0 data. The AERONET program provides a long-term, continuous, and readily accessible public domain database of aerosol optical properties. The network imposes standardization of instruments, calibration, processing, and distribution. The location of individual sites is available from the AERONET web site http://aeronet.gsfc.nasa.gov/.

Lary et al. (2009) first identified all MODIS overpasses of the AERONET sites throughout the lifetime of the two MODIS missions. Then used the single green band MODIS AOD (550 nm) in the geographic grid point that contains the AERONET site. AERONET AOD measurements within 30 min of the MODIS observation are averaged. AERONET data are interpolated (in log–log space) to the green band where they are missing. They found a strong correlation between geographic location and bias. For example, there is a negative bias (MODIS underestimation relative to AERONET) over vegetated Western Africa (from Liberia to Nigeria) and a positive bias over the Southwestern U.S. The spatial dependence of the differences between AERONET and MODIS is shown in Fig. 7.

Fig. 7 MODIS bias with respect to AERONET. Computed as a regression with intercept at the origin. Red indicates that MODIS is higher; blue indicates that AERONET is higher. The size of the circle is proportional to the slope of the regression for slope > 1 (where MODIS is higher) and to the inverse of the slope for slope < 1

Machine Learning AOD Bias Correction

Lary et al. (2009) applied two types of machine learning to the correction of the bias between MODIS and AERONET, i.e., neural networks and support vector machines (SVMs). For each of these machine-learning approaches, they used two training data sets, i.e., one for MODIS Aqua and one for MODIS Terra. These training data sets include all contemporaneous measurements of the MODIS instruments and AERONET made from launch to the present that were within 30 min of each other, within a great circle distance of 0.25°, and within a solar zenith angle of 0.1°. For MODIS Aqua, this gave a training record of 7543 points, and for Terra, 13,034 points.

The purpose of training a machine-learning algorithm is to construct a mapping between a set of input variables and an output variable (i.e., a multivariate nonlinear nonparametric fit). For each data set, the inputs were the surface type, the solar zenith angle, the solar azimuth angle, the sensor zenith angle, the sensor azimuth angle, the scattering angle, the reflectance, and the MODIS AOD. For each data set, the output was the AERONET AOD at 550 nm.

Figure 8c and d shows the result of performing a neural network bias correction. We see that the neural network is able to make a substantial improvement in the correlation coefficient with AERONET: an improvement from 0.86 to 0.96 for MODIS Aqua and an improvement from 0.84 to 0.92 for MODIS Terra.

Figure 8e and f shows the result of performing an SVR bias correction. The SVR makes an even greater improvement than the neural network correction, improving the correlation coefficient from 0.86 to 0.99 for MODIS Aqua and from 0.84 to 0.99 for MODIS Terra.

Examining the linear regression on the SVM fit, we see that the intercept (bias) is considerably reduced, from 0.03 to 0.0005 for Aqua and from 0.03 to 0.0001 for Terra. In addition, the slope of the SVM fit is almost 1 (0.99) for both Aqua and Terra.

Overall, the machine-learning results of Lary et al. (2009) (Fig. 8) show us that there is opportunity in the MODIS aerosol algorithm to improve the accuracy of the AOD retrieval, as compared with AERONET, and that this improvement is linked to surface type. We can use information from AERONET, from other satellite sensors such as MISR, and from detailed field experiments to continue to test and refine the assumptions in the MODIS algorithm. The results from the machine-learning analysis that point to surface type as the missing piece of information will allow us to focus the refinement procedure where it will help most.

Machine-learning algorithms were able to effectively adjust the AOD bias seen between the MODIS instruments and AERONET. SVMs performed the best, improving the correlation coefficient between the AERONET AOD and the MODIS AOD from 0.86 to 0.99 for MODIS Aqua and from 0.84 to 0.99 for MODIS Terra. Key in allowing the machine-learning algorithms to “correct” the MODIS bias was provision of the surface type and other ancillary variables that explain the variance between the MODIS and AERONET AOD.

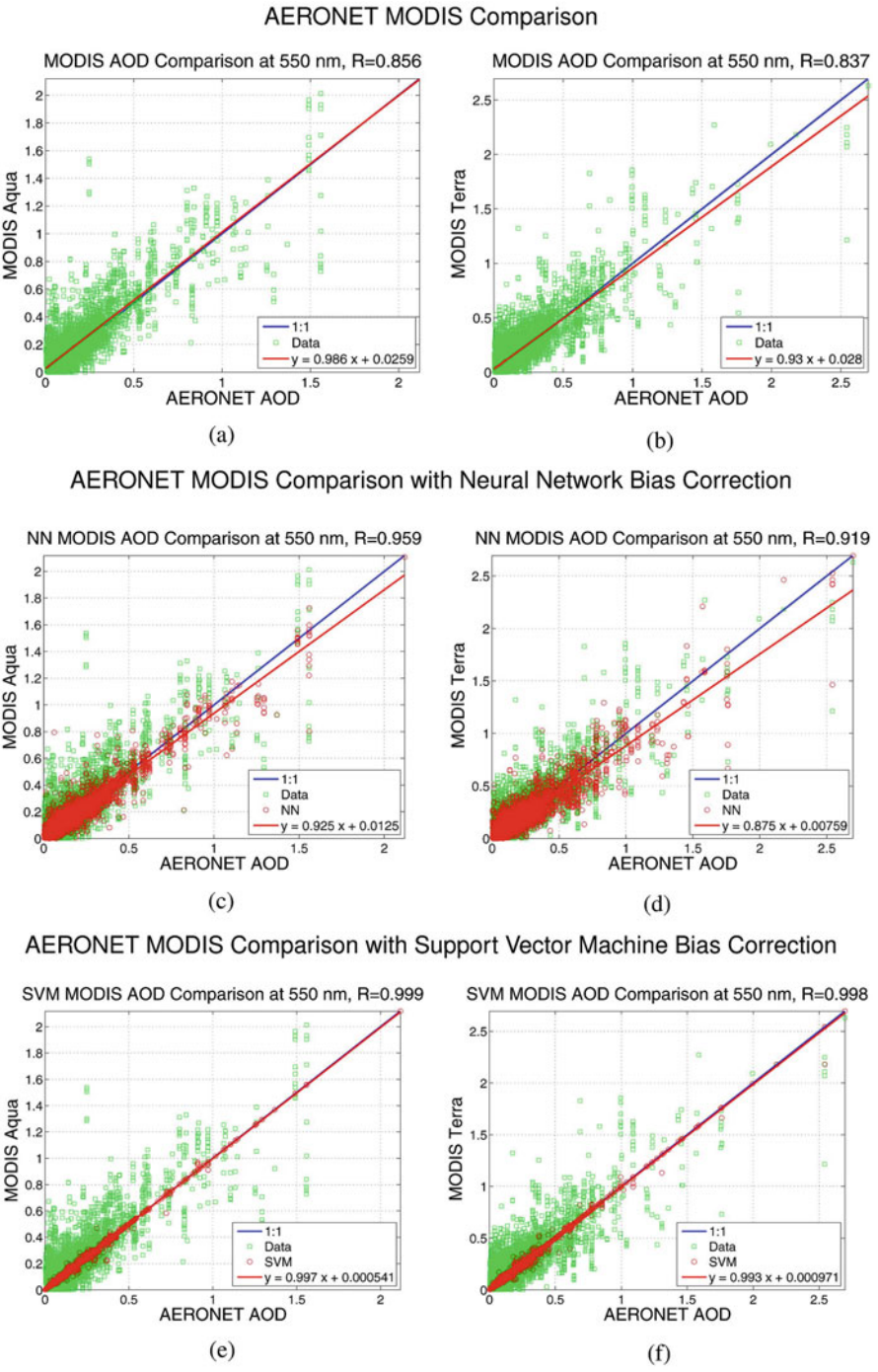

Fig. 8 (continued)

Machine Learning for New Product Creation

Let us know turn our attention to an example of creating new data products through the holistic use of satellite and in situ data. A new data product that is of societal significance.

Airborne Particulates

There is an increasing awareness of the health impacts of particulate matter and a growing need to quantify the spatial and temporal variations of the global abundance of ground level airborne particulate matter ($PM_{2.5}$). In March 2014, the World Health Organization (WHO) released a report that in 2012 alone, a staggering 7 million people died as a result of air pollution exposure (), one in eight of the total global deaths. A major component of this pollution is airborne particulate matter (e.g., $PM_{2.5}$ and PM_{10}).

The recent study by Lary et al. (2014) used machine learning to provide daily global estimates of airborne $PM_{2.5}$ from 1997 to 2014. This was achieved utilizing by using a massive amount of data (40 TB) from a suite of about 100 remote sensing and meteorological data products together with ground based observations of $PM_{2.5}$ from 8329 measurement sites in 55 countries taken between 1997 and 2014. This data was used to train a machine learning algorithm to estimate the daily distributions of $PM_{2.5}$ from 1997 to 2014. This allowed the creation of a new global $PM_{2.5}$ product at 10 km resolution from August 1997 to present (Lary et al., 2014). This new dataset is specifically designed to support health impact studies. Lary et al. (2014) showed some examples of this global $PM_{2.5}$ dataset and finish by examining a mental health Emergency Room admissions in Baltimore, MD. They demonstrate that the new $PM_{2.5}$ data product can reliably represent global observations of $PM_{2.5}$ for epidemiological studies. They showed that airborne particulates can have some surprising associations with health outcomes. As an example of this, Lary

Fig. 8 Scatter diagram comparisons of AOD from AERONET (*x*-axis) and MODIS (*y*-axis) as green circles overlaid with the ideal case of perfect agreement (blue line). The measurements shown in the comparison were made within half an hour of each other, with a great circle separation of less than 0.25° and with a solar zenith angle difference of less than 0.1°. The left-hand column of plots is for MODIS Aqua, and the right-hand column of plots is for MODIS Terra. The first row shows the comparisons between AERONET and MODIS for the entire period of overlap between the MODIS and AERONET instruments from the launch of the MODIS instrument to the present. The second row shows the same comparison overlaid with the neural network correction as red circles. We note that the neural network bias correction makes a substantial improvement in the correlation coefficient with AERONET. An improvement from 0.86 to 0.96 for MODIS Aqua and an improvement from 0.84 to 0.92 for MODIS Terra. The third row shows the comparison overlaid with the SVR correction as red circles. We note that the SVR bias correction makes an even greater improvement in the correlation coefficient than the neural network correction. An improvement from 0.86 to 0.99 for MODIS Aqua and an improvement from 0.84 to 0.99 for MODIS Terra

et al. (2014) presented an analysis of Baltimore schizophrenia Emergency Room admissions in the context of the levels of ambient pollution. $PM_{2.5}$ had a statistically significant association with some aspects of mental health.

A useful validation of the new $PM_{2.5}$ data product is to survey the key features of the global $PM_{2.5}$ distribution and see if they capture what we expect to find and what has been reported in the literature. In Fig. 9a we see that the eastern half of the USA has a higher average abundance of $PM_{2.5}$ than the western half of the USA with the exception of California. This is consistent with the overlaid EPA observations shown as color filled circles. The color fill for the observations uses the same color scale as the machine learning estimate depicted using the background colors. There are persistently high levels of $PM_{2.5}$ in Mexico's dusty and desolate Baja California Sur. The particularly high values are in Mulegé Municipality close to Guerrero Negro (marked A in panel (a) of Fig. 9). Straddling the region close to the Mexico, Arizona, and California borders is the Sonoran Desert. This is a region characterized by a high average $PM_{2.5}$ abundance (marked B) and haboobs, massive dust storms. The Sonoran desert has an area of 311,000 square kilometers and is one of the hottest and dustiest parts of North America. This is clearly evident in the high 16-year average $PM_{2.5}$ abundance in this region. The persistently high $PM_{2.5}$ abundance associated with Los Angeles is visible (marked C). The regions of high population density usually coincide with the region of high particulate abundance. California's heavily agricultural Central Valley has a high $PM_{2.5}$ loading (marked D), note the good agreement of our estimates with the 16 year average observations. The EPA has designated Central Valley as a non-attainment area for the 24-h $PM_{2.5}$ National Ambient Air Quality Standards (NAAQS). The high $PM_{2.5}$ abundance associated with the Great Salt Lake Desert in northern Utah close to the Nevada border is clearly visible (marked E). There is a nearby measurement supersite at Salt Lake City recording a particulate abundances consistent with our estimates. Mexico City is known for its high levels of particulates and is clearly visible (marked F) as a localized hot spot. Close to the Mexico/Texas border we see the elevated $PM_{2.5}$ abundance associated with the Chihuahuan Desert and the Big Bend Desert (marked G). Dust storms in this area often impact El Paso in Texas and Ciudad Juarez in Mexico. The Ohio River Valley (marked H) encompasses several states and is home to numerous coal-fired power plants, chemical plants, and industrial facilities, leading to high levels of ambient particulates. The Ohio River Valley has a higher average abundance of $PM_{2.5}$ than the rest of the East Coast. Our analysis agrees closely with the in-situ observations for the Athens super-site. The Piura desert in Northern Peru (marked I) on the coast and western slopes of the Andes is a region of high particulate abundances. The region in South America from the high Andean semi-arid Altiplano basin in the north, coming down through the Salar de Uyuni Desert (the world's largest salt flats), passing by Santiago in Chile and San Miguel de Tucumn, San Juan and Mendoza in Argentina, and down to the Neuquén Basin in the south is characterized by a high abundance of particulates from a combination of dust, salt, and pollution (marked J). The southern Amazon in Bolivia and the surrounding region has a lot of burning leading to persistently high particulate abundances (labeled K).

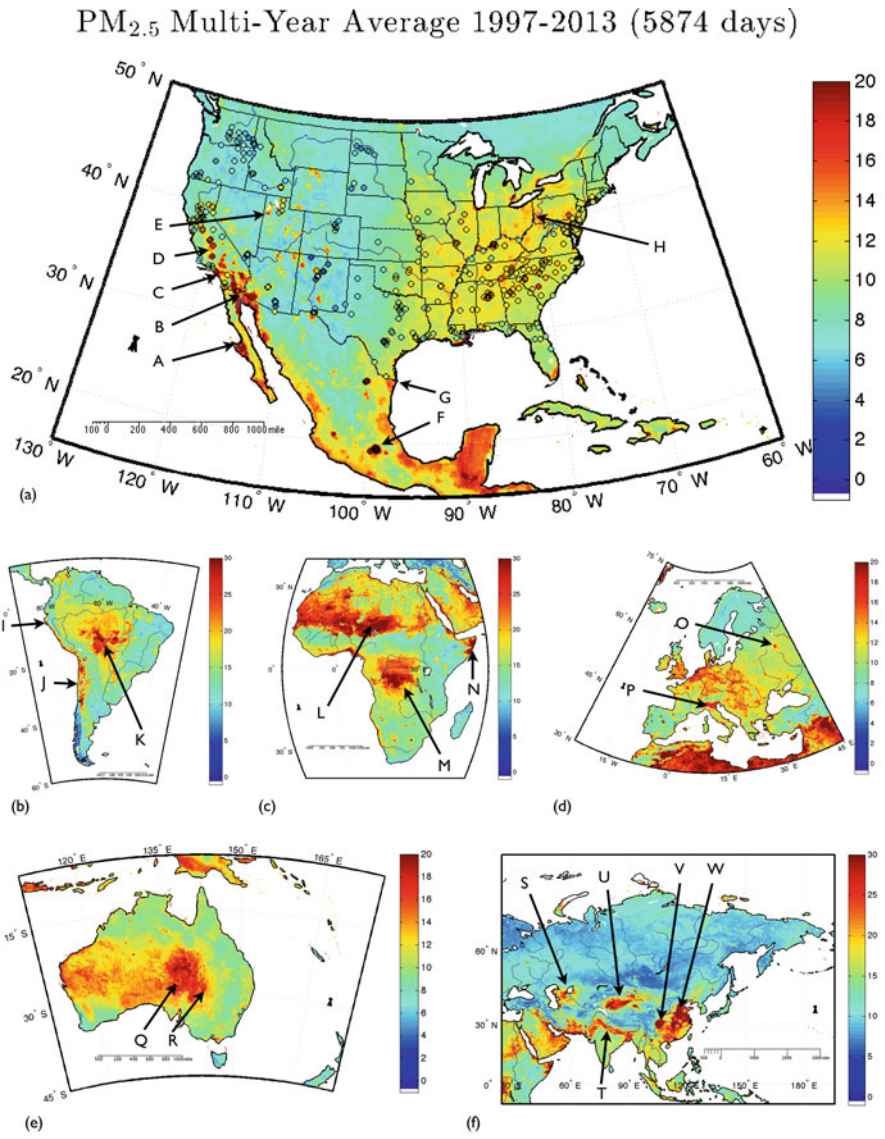

Fig. 9 The average of the estimated surface $PM_{2.5}$ abundance of the 5874 daily estimates from August 1, 1997, to August 31, 2013 in $\mu g/m^3$ for (**a**) the USA, (**b**) South America, (**c**) Africa, (**d**) Europe, (**e**) Australia, and (**f**) Asia

The Bodélé depression is Chad's lowest point on the Sahara's southern edge that supplies the Amazon forest with the majority of its mineral dust. The high abundance of $PM_{2.5}$ over the Bodélé is clearly visible (marked L). Typically there are dust storms originating from the Bodélé depression on around 100 days a year.

The low flat desert in the North African Western Sahara is some of the most inhospitable and arid land on earth and a substantial dust source, clearly visible in the high abundance of $PM_{2.5}$. Burning in the Democratic Republic of the Congo (marked M) leads to high levels of particulates. Much of coastal Somalia is desert characterized by high levels of particulates (marked N).

The Italian Po valley (marked P in Fig. 9) has some of the highest average abundance of particulates in Europe. Industrial emissions coupled with persistent fog leading to heavy smog. High levels of $PM_{2.5}$ are found in the Netherlands and North-west Germany. An example of a local pollution hotspot in Europe is Moscow (marked O).

Lake Eyre is Australia's largest lake and lowest point (marked Q). When the lake has dried out a salt crust remains. When Lake Eyre is dry it is typically Australia's largest dust source, Lake Eyre usually only fills with water after the heavy rains that typically occur once every 3 years, during these periods the $PM_{2.5}$ abundance in the vicinity of Lake Eyre is lower than usual. Just east of the Lake Eyre Basin is the Strzelecki Desert another major Australian dust source (marked R). The arid region just south of the Hamersley Range in Western Australia, the Gibson Desert, Great Victoria Desert and MacDonnell Ranges are also dusty environments with elevated average abundances of $PM_{2.5}$.

Asia has some of the highest particulate abundances anywhere on earth. The Aral Sea (marked S) lying across the border of Kazakhstan and Uzbekistan is heavily polluted with major public health problems. The Ganges Valley is home to 100 million people and is highly polluted (marked T). The cold Taklimakan Desert of northwest China is a major source of $PM_{2.5}$ (marked U). Particularly high levels of particulates are found in the Sichuan Basin (marked V) and in western China in the region from Beijing in the North down to Guangxi in the south (marked W).

Tracer Correlations

The spatial distributions of atmospheric trace constituents are in general dependent on both chemistry and transport. Compact correlations between long-lived species are well observed features in the middle atmosphere. The correlations exist for all long-lived tracers—not just those which are chemically related—due to their transport by the general circulation of the atmosphere. The tight relationships between different constituents have led to many analyses where measurements of one tracer are used to infer the abundance of another tracer. These correlations can also be used as a diagnostic of mixing and to distinguish between air-parcels of different origins.

Of special interest are the so-called "long-lived tracers": constituents such as nitrous oxide (N_2O), methane (CH_4), and the chlorofluorocarbons (CFCs) that have long lifetimes (many years) in the troposphere and lower stratosphere, but are destroyed rapidly in the middle and upper stratosphere.

The correlations are spatially and temporally dependent. For example, there is a compact-relation regime in the lower part of the stratosphere and an

altitude-dependent regime above this. In the compact-relation region, the abundance of one tracer is uniquely determined by the value of the other tracer, without regard to other variables such as latitude or altitude. In the altitude-dependent regime, the correlation generally shows significant variation with altitude.

The description of such spatially and temporally dependent correlations is usually achieved by a family of correlations. However, a single neural network is a natural and effective alternative.

Reconstructing CH_4–N_2O Correlations

The motivation for this study was preparation for a long-term chemical assimilation of Upper Atmosphere Research Satellite (UARS) data starting in 1991. For this period we have continuous version 19 data from the Halogen Occultation Experiment (HALOE) but not observations of N_2O as both ISAMS and CLAES failed. In addition we would like to constrain the total amount of reactive nitrogen, chlorine, and bromine in a self-consistent way (i.e., the correlations between the long-lived tracers are preserved). Tracer correlations provide a means to do this by using HALOE CH_4 observations.

Figure 10a shows the CH_4–N_2O correlation from the Cambridge 2D model overlaid with a neural network fit to the correlation (Lary et al., 2003). The neural network used was a feed-forward multilayer perceptron. There were four inputs, one output, and one hidden layer with eight nodes. A non-linear activation function was used. The training dataset contained 1292 patterns, sampling the input space completely as shown in Fig. 10. The network was constrained for 106 epochs (iterations).

The correlation coefficient between the actual solution and the neural network solution was 0.9995. Figure 10 panel (b) shows how the median fractional error of the neural network decreases with epoch (iteration). Both CH_4 and pressure are strongly correlated with N_2O as can be seen in panels (c) and (d). Latitude and time are only weakly correlated with N_2O as can be seen in panels (e) and (f). Even though the correlation with time of year and latitude is relatively weak it still does play a role in capturing some of the details of the CH_4–N_2O correlation in Panel (a).

A polynomial or other fit will typically do a good job of describing the CH_4–N_2O correlation for high values of CH_4 and N_2O. However, for low values of CH_4 and N_2O there is quite a spread in the relationship which a single curve can not describe. This is the altitude dependent regime where the correlation shows significant variation with altitude.

Figure 10c shows a more conventional fit using a Chebyshev polynomial of order 20. This fit was chosen as giving the best agreement to the CH_4–N_2O correlation after performing fits using 3667 different equations. Even though this is a good fit the spread of values cannot be described by a single curve. However, a neural network trained with the latitude, pressure, time of year, and CH_4 volume mixing ratio (v.m.r.) (four inputs) is able to well reproduce the N_2O v.m.r. (one output), including the spread for low values of CH_4 and N_2O.

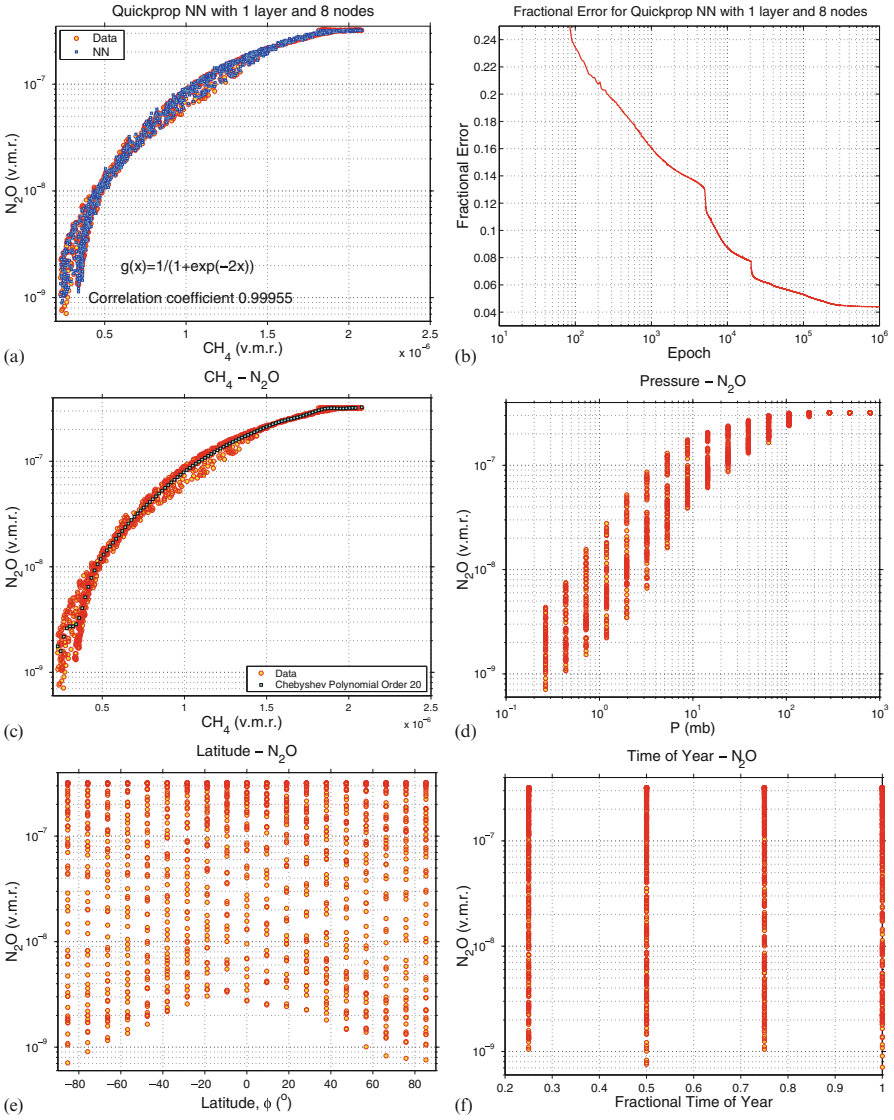

Fig. 10 The neural network used to produce the CH_4-N_2O correlation in Panel (**a**) used a neural network with one hidden layer with eight nodes. The correlation coefficient between the actual solution and the neural network solution was 0.9995. Panel (**b**) shows how the median fractional error of the neural network decreases with epoch (iteration). Both CH_4 and pressure are strongly correlated with N_2O as can be seen in panels (**c**) and (**d**). Latitude and time are only weakly correlated with N_2O as can be seen in panels (**e**) and (**f**). Even though the correlation with time of year and latitude is relatively weak it still does play a role in capturing some of the details of the CH_4-N_2O correlation in Panel (**a**)

Variable scaling often allows neural networks to achieve better results. In this case all variables were scaled to vary between zero and one. If the initial range of values was more than an order of magnitude, then log scaling was also applied. In the case of time of year the sine of the fractional time of year was used to avoid a step discontinuity at the start of the year.

Neural networks are clearly ideally suited to describe the spatial and temporal dependence of tracer-tracer correlations (Lary et al., 2003). Even in regions when the correlations are less compact. Useful insight can be gained into the relative roles of the input variables from visualizing the network weight assignment.

Pollen Estimation

Pollen is known to be a trigger for allergic diseases, e.g. asthma, hay fever, and allergic rhinitis (Oswalt and Marshall, 2008; Howard and Levetin, 2014). It is interesting that a variety of non-respiratory issues such as strokes (Low et al., 2006), and surprisingly, even suicide and attempted suicide (Matheson et al., 2008) have an association with the daily concentration of atmospheric particulates. However, so far, there is no defined threshold amount of pollen known to trigger allergy for sensitive individuals (Voukantsis et al., 2010). One of the factors for the lack of knowledge of the threshold amount of pollen is the absence of an accurate estimation on a fine spatial scale of the hourly, bi-hourly, or daily amount of pollen. Individual physiological differences such as gender and age among sensitive people also adversely affect in knowing the threshold amount of pollen in the surrounding (Britton et al., 1994; Ernst et al., 2002).

Of all plants, weeds, and particularly those of the *Ambrosia* species, e.g. *Ambrosia artemisiifolia* (common ragweed), *Ambrosia trifida* (giant ragweed) are major producers of large amounts of pollen. For example, a common ragweed can produce up to about 2.5 billion pollen grains per plant per day (Laaidi et al., 2003). *Ambrosia artemisiifolia* and *Ambrosia trifida* combined can produce more allergens than all other plants combined (Lewis et al., 1983). Grasses (e.g., *rye grass*) are also known to trigger an allergic response. Following *Ambrosia artemisiifolia*, grass pollen are known for their high allergic potency than most weeds (Esch et al., 2001; Lewis et al., 1983). Tree pollen can cause an allergic response, but one that is typically less than that of weeds and grasses, although in some regions tree pollen can trigger a significant allergic response. For instance, the airborne concentration of Mountain cedar pollen grains can reach tens of thousand of pollen grains per cubic meter and trigger a significant allergic response in central Texas during winter, known as cedar fever (Andrews et al., 2013; Ramirez, 1984).

Both global climate change and air pollution affect the abundance of airborne pollen, and consequently, its allergic impact (Kinney, 2008; Wayne et al., 2002; Voukantsis et al., 2010). For example, the abundance of pollutants such as CO_2, Wayne et al. (2002) and NO_2 (Zhao et al., 2016) can affect the extent of growing season of major pollen producing plants, and thereby also affect the airborne pollen concentration as well as altering the onset and end dates of seasonal allergies.

Overall, more people are exposed to pollen and sensitive individuals become exposed to large amount of pollen for longer period of time over larger areas.

Globally millions of people are affected by seasonal allergies, and the number of people affected is increasing each year. In North American alone, as of 2008, about 50 million adult Americans and 9% of children aged below 18 have experienced pollen caused allergies (Howard and Levetin, 2014). Similarly, in Europe about 15 million people are affected by hay fever, asthma, and rhinitis (D'amato and Spieksma, 1991). Hence, pollen allergies are becoming an increasingly significant environmental health issue. Hence, just as accurate daily weather forecasts are of significant use, accurate daily pollen forecasts are likely to become increasingly important.

Remote Sensing has been employed to study atmospheric pollen concentrations. For example, the polarization of LIDARs has been used to observe the airborne tree pollen abundance at Fairbanks Alaska (Sassen, 2008). In this case, the pollen produces a depolarization of the LIDAR backscattering signals from the lower atmosphere. The light scattering properties of pollen are also manifested in the shape of the solar corona they create. The shape of the solar corona associated with pollen depends on the shape of the pollen grains and their atmospheric concentration (Tränkle and Mielke, 1994). However, this approach can be complicated as atmospheric light scattering is also caused by other airborne particulates.

Common pollen estimation techniques, particularly those made in Europe, stress the importance of meteorologic variables (Kasprzyk, 2008). Usually forecasting the amount of airborne pollen is based on the interaction of atmospheric weather and pollen (Arizmendi et al., 1993). Meteorologic variables such as the daily mean, maximum, change in temperature and dew point variables show positive correlation with the pollen concentration (Kasprzyk, 2008). Kasprzyk (2008) found that atmospheric humidity shows negative correlation to the pollen concentration. Other studies show that temperature, precipitation, and wind speed are significant meteorologic parameters in estimating pollen concentration (Stark et al., 1997).

Most of these meteorologic variable based forecasting methods employed statistical methods such as linear regression, the polynomial method and time serious analysis (Sánchez-Mesa et al., 2002). Only few studies used advanced machine learning methods such as neural network (Sánchez-Mesa et al., 2002; Rodríguez-Rajo et al., 2010; Puc, 2012; Voukantsis et al., 2010) and random forest (Nowosad, 2016) for pollen forecasting and support vector machines are applied for related environmental studies (Voukantsis et al., 2010; Osowski and Garanty, 2007).

Predicting Pollen Abundance

Over the past decade neural networks have been applied to study pollen of different species over the European region. For example, Csépe et al. (2014) used different Computational Intelligence (CI) methods to predict the *Ambrosia* pollen at two different places in Hungary and France. Castellano-Méndez et al. (2005) and Puc (2012) have employed the neural network to predict *Betula* pollen over Spain and

Table 1 Name and type of predictors (input variables) used for our machine learning training. Parameters consist of environmental and NEXRAD radar measurements

Parameter	Unit	Type
Vegetation greenness fraction	Fraction	Env.
Leaf area index	m^2	Env.
Roughness length, sensible heat	m	Env.
Displacement height	m	Env.
Energy stored in land	Jm^{-2}	Env.
Mean reflectivity	dB	NEXRAD
Mean Doppler velocity	ms^{-1}	NEXRAD
Mean spectral width	ms^{-1}	NEXRAD
Reflectivity [10–10 dB]	dB	NEXRAD
Velocity [10–10 dB]	ms^{-1}	NEXRAD
Spectral width [10–10] dB	ms^{-1}	NEXRAD
Reflectivity [20–20 dB]	dB	NEXRAD
Velocity [20–20 dB]	ms^{-1}	NEXRAD
Spectral width [20–20 dB]	ms^{-1}	NEXRAD
Reflectivity [40–40 dB]	dB	NEXRAD
Velocity [40–40 dB]	ms^{-1}	NEXRAD
Spectral width [40–40 dB]	ms^{-1}	NEXRAD
Wind direction at altitude 50 m	Degree	NEXRAD
Wind speed at altitude 50 m	ms^{-1}	NEXRAD

Poland, respectively. Recently, Nowosad (2016) used the random forest method to forecast different tree pollen species.

In this study we used random forests, neural networks, and support vector machines to estimate daily *Ambrosia* pollen concentration at Tulsa, Oklahoma (location: 36.1511°N, 95.9446°W). We used a combination of environmental parameters and NEXRAD radar measurements. The combined parameters are listed in Table 1. The daily pollen concentration used in the training of our machine learning algorithms was obtained using a Burkhard spore trap at the University of Tulsa, Oklahoma.

After pollen is produced in the plant anthers its emission dispersion and deposition is influenced by meteorological variables such as the temperature, wind speed, and direction and pressure (Kasprzyk, 2008; Csépe et al., 2014; Howard and Levetin, 2014). Other meteorologic parameters such as dew point, humidity, rainfall sunshine duration are also known to affect pollen emission and distribution (Kasprzyk, 2008).

We used a set of environmental and NEXRAD radar parameters (Table 1) in our machine learning training. Environmental parameters such as vegetation greenness fraction, roughness length (sensible heat), energy stored in all land reservoirs, and displacement height and leaf area index are selected. The other set of data we used are the NEXRAD measurements which consist of the reflectivity, Doppler velocity, and spectral width which represent, respectively, the amount of back scattered signals from a scattering volume, the velocity of the scatterer along the radar line of sight and the width of the power spectrum. All NEXRAD measurements are taken

at the lowest elevation. Additionally the NEXRAD provides measurements of the vertical profile of the direction and speed of the wind from about near the surface of the Earth. The dual polarization measurements: differential reflectivity, differential phase, and correlation coefficient use the horizontal and vertical polarization signals and are particularly suited for particle identification. In this study we do not use the dual polarization (polarimetric) NEXTRAD measurements as we have only few days of the measurements in contrary to the ideal high dimensional data requirement for machine learning.

The three machine learning methods were trained on the entire data set to assess their performance in predicting the *Ambrosia* pollen. The scatter diagrams are shown in Fig. 11. We also used a Newton-Raphson based recursive Random Forest technique that has been developed in order to improve the accuracy. The method includes error estimation and correction. In order to evaluate the performance of the machine learning methods independently, 10% of the data are randomly selected and withdrawn for validation from the training and the remaining 90% of the data is then used for training the model. After developing the model, its performance is tested using the independent 10% of the validation dataset that was not used in training the machine learning regression. These results are shown in Fig. 11. Panels (a)–(c) in Fig. 11 show scatter plots of prediction made by the support vector machine, neural network, and random forest machine learning methods, respectively, using

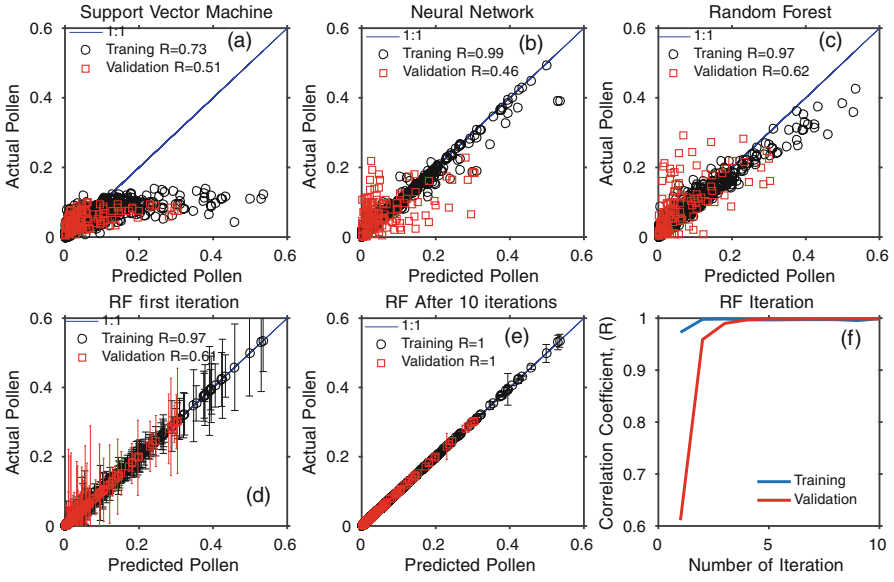

Fig. 11 Showing scatter plots of actual and predicted pollen for the support vector machine (panel **a**), neural network (panel **b**), random forest (panel **c**). Panel (**d**)–(**f**) shows results of an iteration method applied to the random forest. Panel (**d**) and (**e**) shows results of the first iteration 10th iterations for the training (black circles) and validation data (red squares). Panel (**f**) depicts plots of the correlation coefficient for the training and validation data versus iteration number

the training data (black circles) and the validation data (red squares). Results of the iterative method applied to the random forest method are given by panels (d)–(f). The random forest machine learning is trained using 200 decision trees.

From the top three panels of Fig. 11, we observe that the neural network and random forest methods produced better predictions than the Support Vector machine. The random forest method produced the best independent validation results (correlation coefficient, 0.62) of all the three methods. The high correlation value of neural network found using the training data (correlation coefficient 0.99) is not reproduced in the independent validation test which had a correlation coefficient of only 0.46. Error bar plots for the training and the validation data for the first iteration of the random forest are given by panel (d) in Fig. 11. We see that predictions using both the training and validation data exhibit large errors and a low correlation coefficient. Interestingly, after a few iterations the random forest produced results with significantly reduced errors and correlation values close to 1 (panel (e) in Fig. 11). Panel (f) in Fig. 11 shows the correlation coefficient values between the normalized estimated and actual pollen for the training (blue curve) and validation data (red curve) sets for 10 iterations. We observe that the iterative of the random forest method has reduced the error significantly and the correlation coefficient values converge to one for both training and validation data sets.

The upper panel of Fig. 12 shows a comparison of the actual and predicted pollen using the recursive random forest. Another important application of machine learning methods is the selection of the best features (variables) that contribute most to the prediction and ranking them in order of the importance. In this way we can determine the most important predictor variables and estimate the output leaving features that contribute less. The random forest provides such a ranking based on criteria attributed to the splitting variable in the data sampling to form decision tree (Genuer et al., 2010; Kotsiantis et al., 2007; Friedman et al., 2001).

The lower panel of Fig. 12 shows the ranking of the relative importance of the variables provided by the random forest with 200 trees. The most important factors in estimating the pollen were the leaf area index, vegetation greenness function, and displacement height.

Using Machine Learning for Ocean Data Products

Let us take a look at using machine learning to estimate chromophoric dissolved organic material (CDOM) absorption. A quality control database has recently been assembled (Aurin and Mannino, 2012) of optical and biogeochemical parameters suitable for supporting the development of ocean color algorithms to retrieve chromophoric dissolved organic material (CDOM) absorption (a_g), CDOM spectral slope (S_g), and Dissolved Organic Carbon (DOC) from satellite-derived, multi-spectral remote sensing reflection (Aurin and Mannino, 2012). This database of sea-surface CDOM is considerably larger (18,035 stations) than NOMADv2 (a subset of SeaBASS that is often used in algorithm development; 1182 stations)

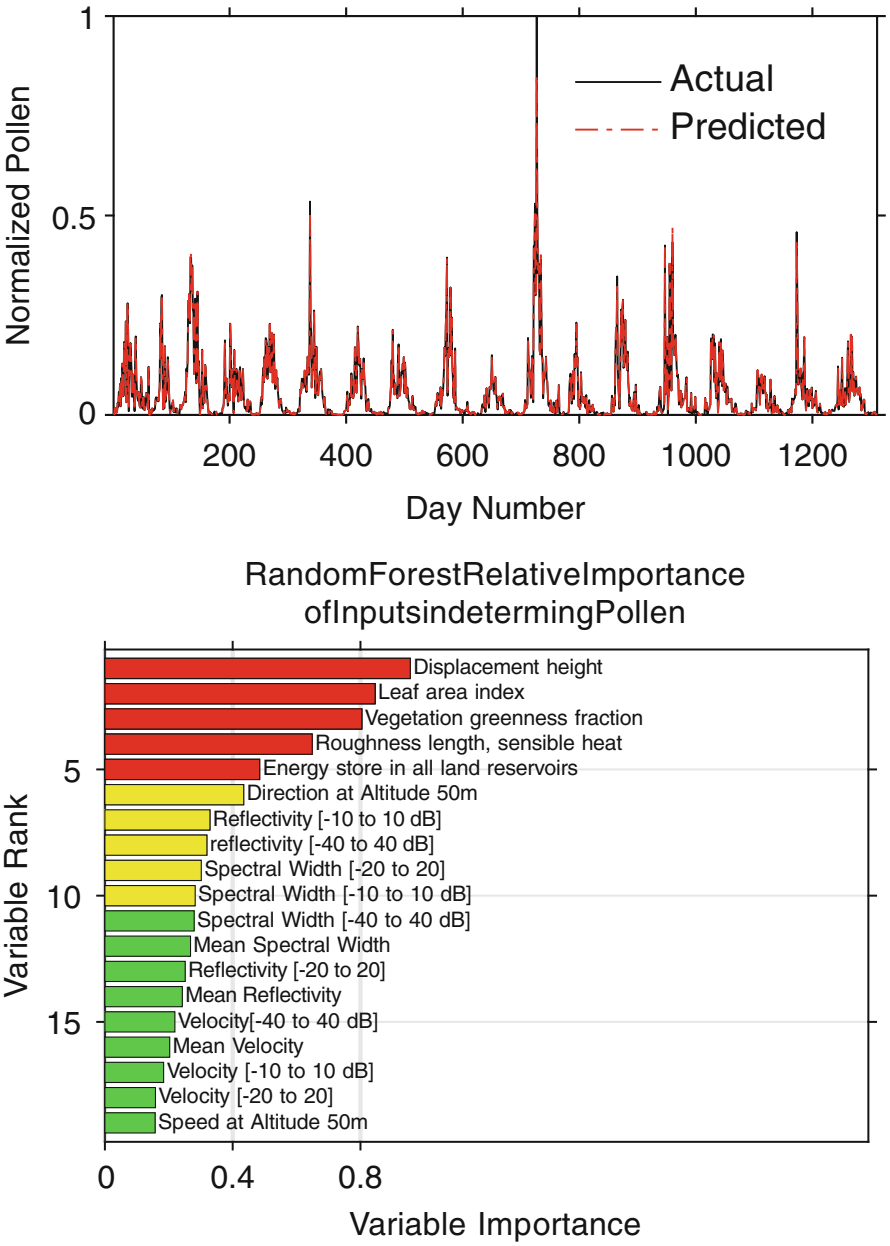

Fig. 12 The upper panel shows the comparison of actual and predicted pollen time series for Tulsa, OK. The lower panel shows the ranking of the relative importance of the variables provided by the random forest with 200 trees

with valid satellite matches at 8654 stations. In the process, estimation of S_g at multiple wavebands between 245 and 715 nm was found to be insensitive to the spectral resolution of a_g, but did depend on reference waveband selection. Global distributions of a_g and S_g are analyzed for patterns related to source, composition, and photo-degradation of CDOM.

CDOM absorbs light most strongly in the ultraviolet and blue region of the spectrum. Non-turbid waters with no CDOM appear blue and as the CDOM concentration increases the color of the water will change from green, through yellow-green to brown. The increasing levels of CDOM diminish light penetration and hence affect the biological activity of aquatic systems by limiting photosynthesis and inhibiting the growth of phytoplankton. CDOM also helps protect organisms from DNA damage by absorbing the harmful UVA and UVB radiation.

Figure 14 shows the global CDOM estimate obtained when the Random Forest multivariate non-parametric fit is used together with a global 9 km SeaWIFS Mapped Seasonal multi-wavelength radiances. The CDOM distribution shown in Fig. 14 is realistic, with low CDOM in regions such as the South Atlantic and South Pacific Gyres, and high CDOM in freshwater and in coastal areas with river outflows.

CDOM abundance is variable, typically with the lower abundances in the open ocean and higher abundances in fresh waters and estuaries. The database (Aurin and Mannino, 2012) was used together with machine learning to develop a set of algorithms to estimate a_g and S_g based on the multi-wavelength remote sensing reflectance observed by SeaWIFS, MODIS Aqua, and MODIS Terra.

Inputs The training dataset used provides training examples to learn from of the multi-wavelength radiances observed by the satellite and of the corresponding ocean parameters such as CDOM. For SeaWIFS the wavelengths used are 412, 443, 490, 510, 555, and 670 nm. For MODIS Terra and Aqua the wavelengths used are 412, 443, 488, 531, 547, and 667 nm. In this study we only consider training examples that were made by the satellites at zenith angles of less than 60°, within 3 h of the in-situ observations, and with a coefficient of variability of less than 0.15.

Fit Quality Figure 13a shows in blue a scatter diagram of the training SeaWIFS CDOM values on the x-axis versus the random forest estimates on the y-axis. The red points show the independent validation dataset. There is a good clustering around the 1:1 line shown in green. The training provides a good fit with an R^2 value of 0.91. The independent verification with the validation dataset has an R^2 value of 0.72 (Table 2). Figure 13b shows the corresponding Taylor diagram, point A is the training observations, point B is the Random Forest fit.

A very useful feature of Random Forests is that they can provide a ranked list of the relative importance of the variables used in producing the regression. This is done by looking at the increase in the mean squared error averaged over all the trees in the ensemble and divided by the standard deviation taken over the trees, for each variable. The larger this value, the more important the variable is in producing a good quality fit. Figure 13c shows the relative importance of the Random Forest inputs. It indicates that in estimating CDOM at 412 nm using SeaWIFS the order of

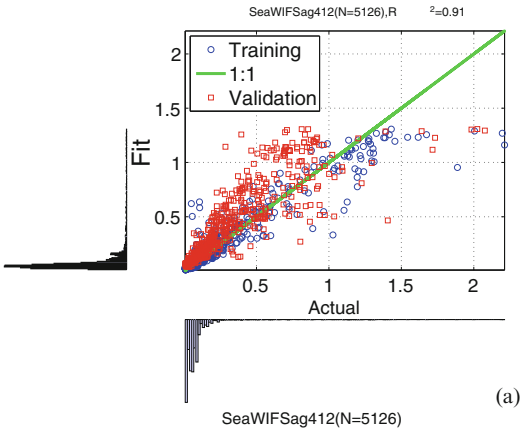

Fig. 13 Panel (**a**) shows in blue a scatter diagram of the training SeaWIFS CDOM values on the *x*-axis versus the estimated values on the *y*-axis. The red points show the independent validation dataset. Panel (**b**) shows the corresponding Taylor diagram, point A is the training observations, point B is the Random Forest fit. Panel (**c**) shows the relative importance of the Random Forest inputs

Table 2 Statistics for the SeaWIFS CDOM training

	σ	ϵ	R^2	n
Training observations	0.15			5126
Training fit	0.13	0.046	0.91	5126
Validation observations	0.33			569
Validation fit	0.35	0.19	0.72	569

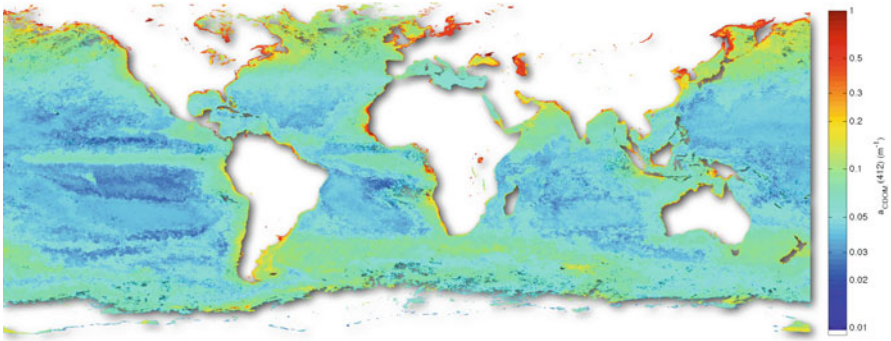

Fig. 14 The global a_{CDOM} (412 nm) estimate obtained when the Random Forest multivariate non-parametric fit is used together with a global 9 km SeaWIFS Mapped Seasonal multi-wavelength radiances

Fig. 15 Dust sources are typically localized point sources

importance of the radiances are 412 nm > 555 nm > 443 nm > 670 nm > 490 nm > 510 nm.

Example Global Distribution Figure 14 shows the global CDOM estimate obtained when the Random Forest multivariate non-parametric fit is used together with a global 9 km SeaWIFS Mapped Seasonal multi-wavelength radiances. The CDOM distribution shown in Fig. 14 is realistic, with low CDOM in regions such as the South Atlantic and South Pacific Gyres, and high CDOM in freshwater and in coastal areas with river outflows.

Dust Source Identification Using Unsupervised Classification

Unsupervised classification can be very useful when we would like to objectively split up our data into different regimes. A good example of this is a study to characterize dust sources (e.g., Fig. 15) (Lary et al., 2016).

Dust sources of many kinds are found globally. One of the most salient features of dust sources is that they are often very localized. For example, in Figs. 15 and 17 we can clearly see that the source of the dust plumes is best described as an ensemble of many point sources, not broad dust emitting regions. Realistically capturing this very localized nature of dust sources has so far largely eluded automated diagnosis, and consequently, description in global models. Invariably current models describe dust sources as rather large scale features, even when vegetation indices and similar approaches are used. This is in marked contrast to what we consistently see in the satellite imagery across the planet (e.g., Figs. 15 and 17).

Identifying dust sources is a critical yet challenging task for the accurate simulation of atmospheric particulate distributions relevant to air quality and climate change.

We take a new and radically different approach to any previous studies that have sought to identify global dust sources on a routine basis. We demonstrate that this new approach employing machine learning is very effective. The approach uses multi-wavelength spectral reflectivity signatures to characterize land surfaces, naturally paving the way for a new class of algorithm ideally suited to fully exploit the next generation of hyper-spectral instrument. The production of thematic maps, such as those depicting land cover, using an image classification is one of the most common applications of remote sensing. New in our approach is that we can both operate at very high spatial resolution and distinguish between types of dust sources. For example, we can easily distinguish between the edge of salt flats (Fig. 17), dried up wadis or lakes, and agricultural sources to name just three of many examples. The only limiting factor for the resolution is the resolution of the satellite imagery.

We employ machine learning to objectively provide an unsupervised multivariate and non-linear classification into a very large number of surface types (in our demonstration study presented below 1000 classes are used) using multi-spectral satellite data. In other words, we do not impose any a priori assumptions, but rather, we let the data speak for itself as to how we should classify surface types. Self-organizing maps (SOMs) are a data visualization and unsupervised classification technique invented by Professor Teuvo Kohonen that reduce the dimensions of data through the use of self-organizing neural networks.

SOMs help us address the issue that humans simply cannot visualize high dimensional data unaided. The way SOMs go about reducing dimensionality is by producing a feature map, usually with two dimensions, that objectively plots the similarities of the data by grouping similar data items together. SOMs learn to classify input vectors according to how they are grouped in the input space. The SOM learns to recognize neighboring sections of the input space. Thus, SOMs learn both the distribution and topology of the input vectors they are trained on.

This approach allows SOMs to display similarities and reduce the dimensionality. A SOM does not assume a priori a functional form for the analyzed data. A noteworthy enhancement of an SOM over principal component analysis is an SOM's ability to represent non-linear functions or mappings.

The premise being that there are very many types of dust sources, from the diatom rich sediments of the Bodélé depression in Chad, to those at the edge of salt flats in Bolivia and Chile (Fig. 17), to those in the coastal Green Mountains of Libya. Each of these dust sources has distinct physical characteristics, and therefore a distinct reflectance signature. If we are able to identify these signatures, then we can map the temporal and spatial evolution of each of these distinct dust sources. Once we have the surface type classification we then seek to identify which small subset of surface classes correspond to various kinds of dust sources. Once we have identified the signature of a wide variety of dust sources we can precisely pick out these locations globally and how their distribution changes with time. This is particularly useful as dust sources are very localized, and some dust sources have a significant seasonal time evolution. Having a methodology to identify the signature of these small-scale regions is invaluable.

The machine learning approach to dust source identification was first conceived in 2010 to face a very practical challenge that the Navy has in producing real time visibility forecasts. If the standard type of dust sources are used it was found that very poor regional visibility forecasts result. However, the quality of the Navy visibility forecasts drastically improved with an analyst (Annette Walker) manually identifying individual dust sources at the heads of plumes by examining sequences of satellite images such as those shown in Fig. 17 and also the EUMETSAT RGB Composites Dust images available online (http://oiswww.eumetsat.org/IPPS/html/MSG/RGB/DUST/). This methodology is very labor intensive and does not lend itself to easy automation. The first prototype dust sources using the machine learning approach described here were devised specifically to automate the dust source identification and also allow for the accurate diagnosis of the time evolution in the spatial extent of the dust sources. Beyond the applications of accurate dust sources for visibility and air quality forecasts, the radiative forcing (RF) due to dust is a key concept in climate change calculations considered by the IPCC for the quantitative comparison of the strength of different human and natural agents causing climate change. Radiative forcing can be categorized into direct and indirect effects. A significant part of the direct effect is the mechanism by which aerosols scatter and absorb shortwave and longwave radiation, thereby altering the radiative balance of the Earth—atmosphere system. Mineral dust is a major component of global aerosols that exert a significant direct radiative forcing. Mineral dust aerosols are produced both naturally ($\approx$70%) and anthropogenically ($\approx$30%).

Our ultimate goal is to identify all the surface locations on the planet that are dust sources. To do this we use a SOM to classify all the land surface locations into a very large set of n categories. In the examples shown here, $n = 1000$. A small subset of these 1000 categories will be regions that are dust sources. Naturally, there are a variety of distinct types of dust sources (e.g., dry river beds, agricultural sources, edge of salt flats, etc.) that we would like to delineate.

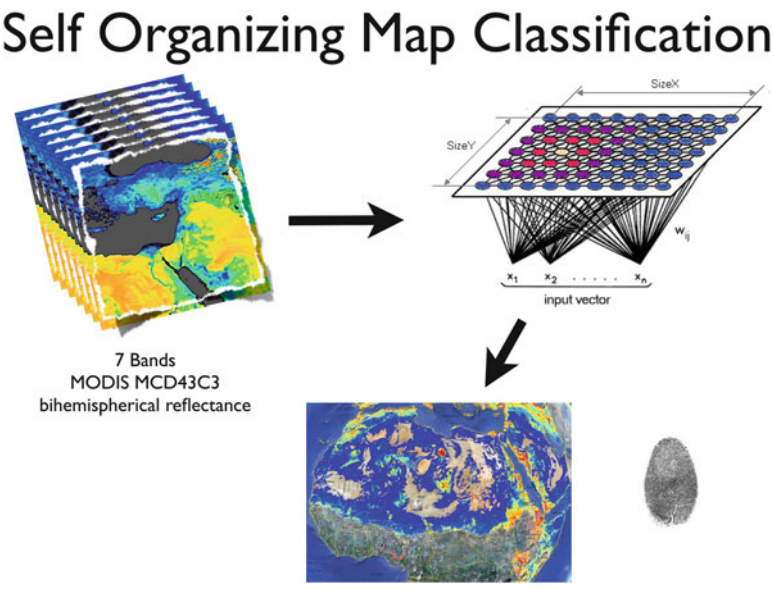

Fig. 16 Schematic of how self-organizing maps have been used in this study to classify land surface pixels into 1000 classes. Then a small subset of these classes are identified as dust sources

To achieve a comprehensive classification we want to consider the conditions present throughout the year, so in the demonstration we took an entire year of the 0.05° resolution MCD43C3 data product (Fig. 16). For this entire year of data, we then calculate the mean, μ for each grid point. This is a massive dataset, and the computational time and memory required to perform the SOM classification increases with the number of data records. For the examples shown here we therefore first restricted our attention to those broad MODIS surface types that may include dust sources, namely: barren or sparsely vegetated surfaces, croplands, grasslands, and open and closed shrublands. These are MODIS surface types 16, 12, 10, 7, and 6, respectively. For each of these surface types we then constructed an input vector that contains seven values, namely for each of the seven bands provided in the MCD43C3 MODIS product the mean, μ, of the directional and bihemispherical reflectance. When training the SOM we use the Euclidean distance to compare the input vectors (each containing seven values).

In order to provide a fine gradation of classification we use the SOM to group together the surface locations into 1000 classes, only a small subset of which correspond to regions that are dust sources. Once the classes that correspond to dust sources have been successfully identified, we have an automated method with which we can identify dust sources that can be routinely executed to provide a regular dust source data product that captures the spatial and temporal evolution of dust sources globally. We utilized the extensive hand classification of very localized dust sources produced by the Navy for the Middle East and South West Asia to guide our initial

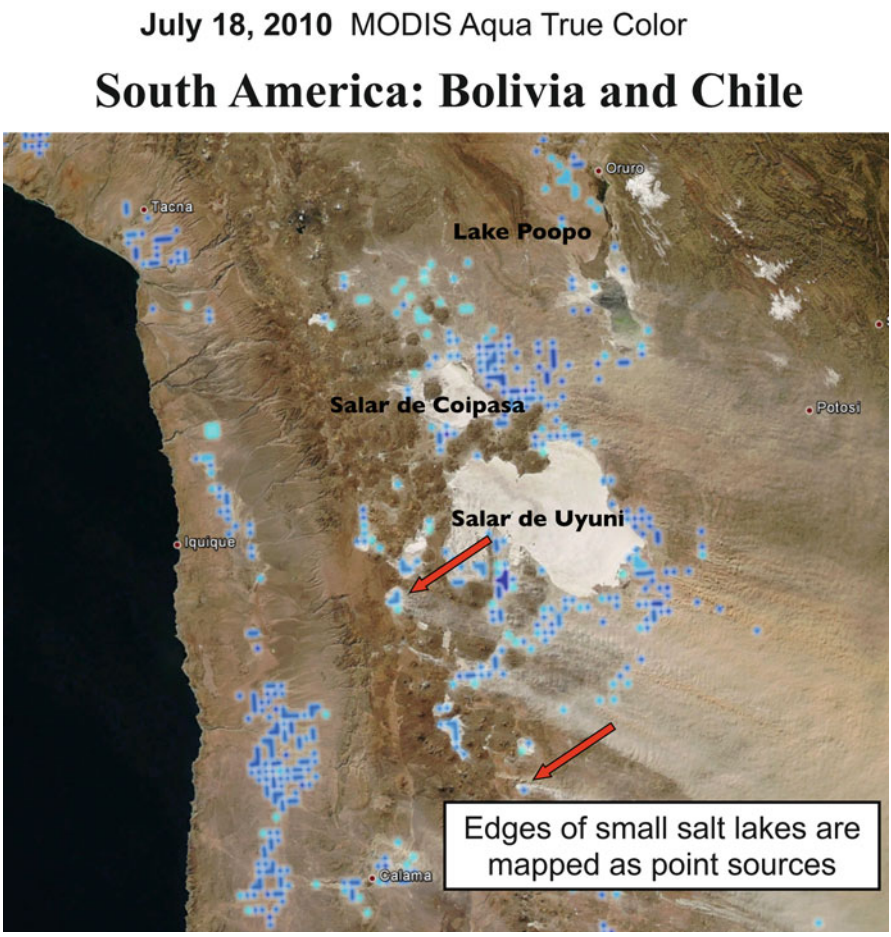

Fig. 17 Example of our machine learning approach correctly identifying very localized point sources around the edge of salt flats in Bolivia and Chile. Notice the narrow dust plumes originating from precisely the identified source regions that have been highlighted in blue and cyan

determination of which of the 1000 classes are dust sources. It is worth noting that the SOM classes are unique and distinct, this will be seen below with the example of the Bodélé Depression. Classes near each other are similar, but distinct.

Bolivia and Chile Salt Flats Dust Event

Figure 17 shows the dust event of July 18, 2010 in the Bolivian Altiplano. This event can be seen clearly in the MODIS Aqua True Color image where dust plumes emanate from fluvio-lacustine deposits and fluviodeltaic sediments around

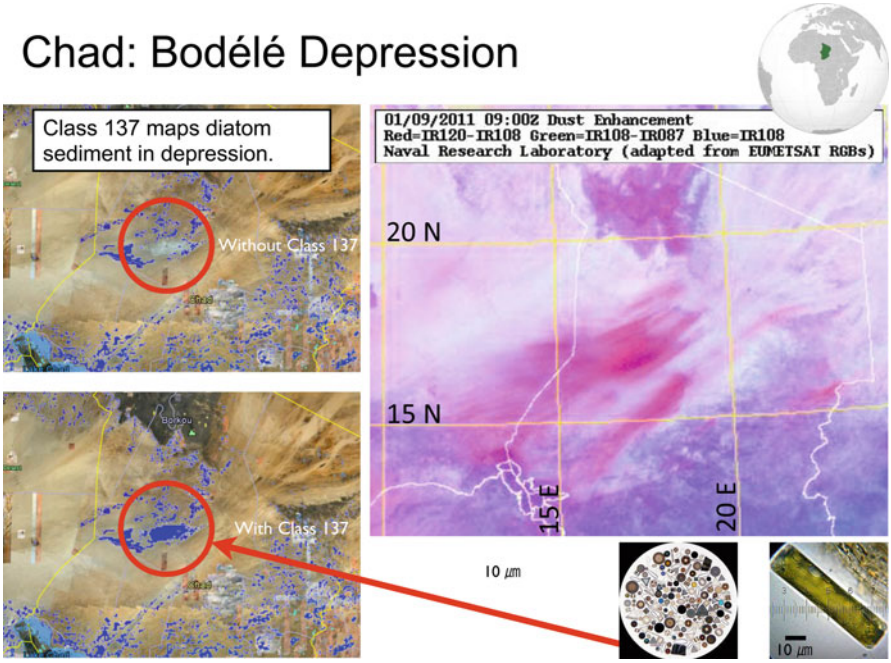

Fig. 18 The Bodélé depression dust event of January 9, 2011 (7Z to 18Z). The right panel shows the NRL processed EUMETSAT MSG/RGB satellite product for January 9, 2011 09Z. The two left panels show the dust sources identified by our approach with (lower) and without (upper) SOM Class 137. Lower right insets show microscopic images of Bodélé diatoms from soils samples taken from the depression

the Salars de Coipasa and Uyuni, Lake Poopo, and other smaller salt flats and lakes. Overlaid are the SOM classes that coincide with active dust sources on the Altiplano. Notice that the salt flats themselves are not dust sources, rather we see the plumes forming around the edges of the flats and lakes. SOMs are very successful in identifying the unique spectral signatures of dust sources. A set of papers which is in preparation will be describing an exhaustive atlas of the global dust sources.

Bodélé Depression Dust Event

Figure 18 shows the Bodélé depression dust event of January 9, 2011 at 09Z that started at 7Z and ceased at 18Z. The Bodélé depression is Chad's lowest point on the Sahara's southern edge. Typically there are dust storms originating from the Bodélé depression on around 100 days a year that supplies the Amazon forest with the majority of its mineral dust (Washington and Todd, 2005; Koren et al., 2006; Washington et al., 2006; Todd et al., 2005; Bouet et al., 2005). The right panel shows the NRL processed EUMETSAT MSG/RGB satellite product. The two left

panels show the dust sources identified by our approach with (lower) and without (upper) SOM class 137. The SOM had automatically determined that the sediment in the Bodélé depression was distinct from the surrounding dust sources and put it in a class all of its own, class 137. Indeed it is different, the Bodélé depression was once filled with a freshwater lake that has long since dried up (Washington et al., 2005). This has left behind diatoms that now make up the surface of the depression. The two key points being, first that the dust source of the Bodélé is distinct from the surrounding dust sources, and second, that it consists of diatoms. This is interesting as if we could devise a way of distinguishing dust sources with containing certain biological materials it would have significant applications for public health issues.

Characterizing Pelagic Habitats Within Coastal Waters

Commercial and recreational fisheries within the Gulf of Mexico contribute significantly to the region's ocean economy making effective management a priority. The goal of fisheries management is to optimize the benefits of living marine resources by addressing threats to a resources' sustainability through conservation, development, and full utilization of the fishery resources to provide food, employment, income, and recreation. Therefore, it is desirable to minimize management actions that may result in negative impacts on fishermen and the coastal community. However, depending on the source of the threat, some management actions such as implementing fishing gear restrictions, time and area closures, and harvest limits, may have direct adverse impacts on fishermen. Often issues impacting fisheries populations arise from degradation or loss of habitat, requiring a different management approach. Coastal and marine habitats can be significantly and rapidly impacted by a number of anthropogenic actions and natural events such as coastal storms, development and hydrological alterations. With approximately 98% of Gulf of Mexico fisheries dependent on estuarine and near-shore habitats at some point in their life cycle (Lellis-Dibble et al., 2008), it is critical that resource managers have the ability to quickly and frequently monitor and assess habitat loss and degradation. However, inconsistencies in the approach that various agencies use for naming habitats make it difficult to develop a region-wide habitat map without standardizing the information.

Unlike most habitat classification systems currently in use, the Coastal and Marine Ecological Classification Standard (CMECS) (Committee et al., 2012) has a water column component which identifies key classifiers required to characterize pelagic habitat types. The vastness and dynamic nature of the ocean's water column limit the feasibility of the frequent in situ sampling that would be necessary to monitor these classifiers and routinely produce region-wide map products. Our ultimate goal is to provide an example of how the Machine Learning classification manifests automatically to the physical classification schemes such as CMECS. Alternatives to in-situ sampling such as remote sensing classification offer a

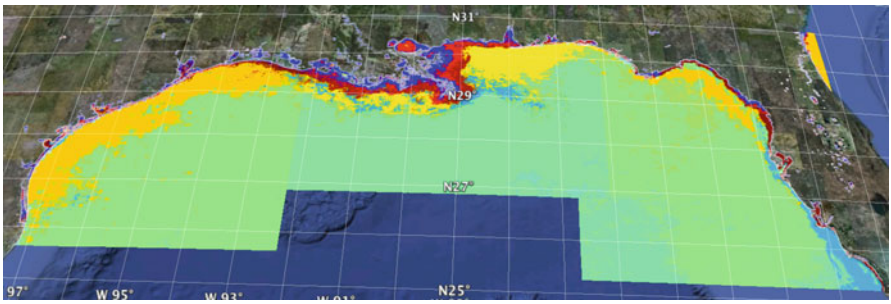

Fig. 19 A self-organizing map classification of the Gulf of Mexico water for 2009. We trained, and applied 5 years of monthly data from 2005 to 2009 using a SOM

proxy for these classifiers, the environmental forcing functions, which shape and determine habitat suitability.

We have used machine learning to objectively provide a water classification of the US Gulf of Mexico using a suite of ocean data products. The unsupervised classification algorithm employed was a self-organizing map (SOM). We used SOM to reduce the dimensionality of the data set. We applied the SOM to identify the multivariate signatures of similar waters and to study the spatial and temporal trends of individual classes over a 5-year period. The input data employed was the sea surface temperature, the chlorophyll concentration, the sea surface salinity, the euphotic depth, and the difference between the sea-bottom and sea-surface temperatures. The output of the analysis is a comprehensive low-dimensional map of US Gulf of Mexico region. This may aid the decision making for the essential habitat zones for conservation and commercial purposes.

The result of the self-organizing map classification is summarized by color-coded maps as shown in Figs. 19 and 21. The SOM essentially group together similar regions as it picks out the characteristic signatures from the data. The similar colors represent the water classified by the method to be on the same class where as the different colors represent the different classification as depicted by the data signatures. Here, the dimension is essentially reduced from 5 to 1. For each SOM class we have computed the spatial extent in square km and how this varies with time. In the time series the annual and biannual signals are clearly evident. A good example of a clear annual cycle in the areal extent is Class 4 in the Eastern Gulf of Mexico. A good example of a clear biannual cycle in the areal extent is Class 61 in the Eastern Gulf of Mexico. Figure 20 shows examples using two SOM classes and the corresponding percentile distributions to the CMECS nomenclatures (Fig. 21).

The dynamics of the region dictates that the characteristics of water should change over the course of the year. In order to study such effects, we computed the areal extent of each of the SOM classes for each month over the 5-year period. By plotting the area of each SOM class as a function of time, we can clearly see the annual and bi-annual cycles in the GOM waters. Characteristic to their SOM class, the areas of the classes change depending upon the time of the year. This gives rise

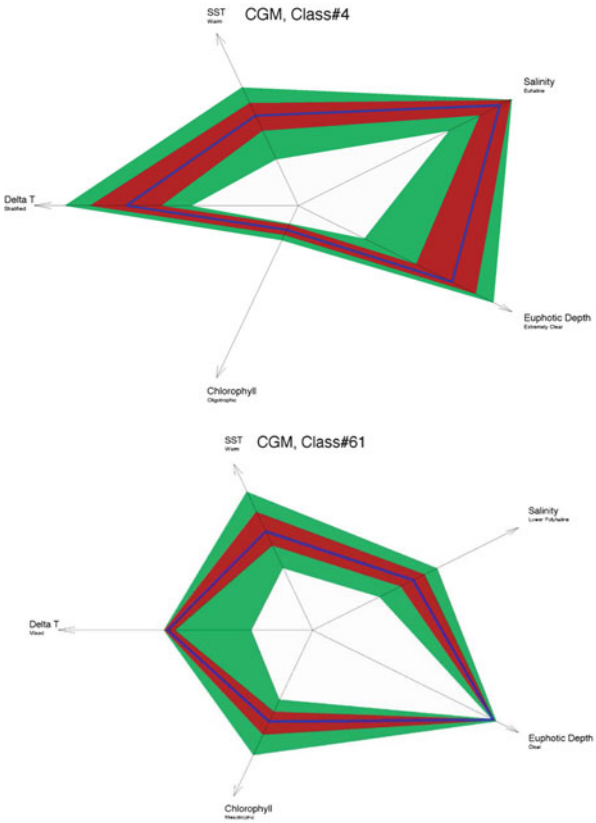

Fig. 20 Webplots showing the categorical variation of data components for the SOM classes 4 and 61. The distributions of the variables are represented by the 5th, 25th, 50th, 75th, and 95th percentile values on the corresponding axes. The upper panel shows the CMECS categories for SOM class 4, associated with waters that have warm sea surface temperatures, are euhaline, extremely clear, oligotrophic and stratified. The lower panel shows the CMECS categories for the SOM class 63, associated with waters that are temperate, mesohaline, moderately clear, mesotrophic, and mixed

to a time series. The pattern repeats itself over the 5-year period. By using Fourier analysis we have identified the frequency distribution of the time series. Figure 22 shows two example SOM classes for the central Gulf of Mexico. We have computed the area of the SOM classes for each month for each class for all regions.

Class 61 lies in the near-shore region and has a clear bi-annual cycle. The areal extent of this class is greatest in April and October, and least in January. Similarly, SOM class 4, which lies in the offshore region, shows a clear annual cycle. The area of class 4 has a maximum during the summer months and has a minimum during the winter months.

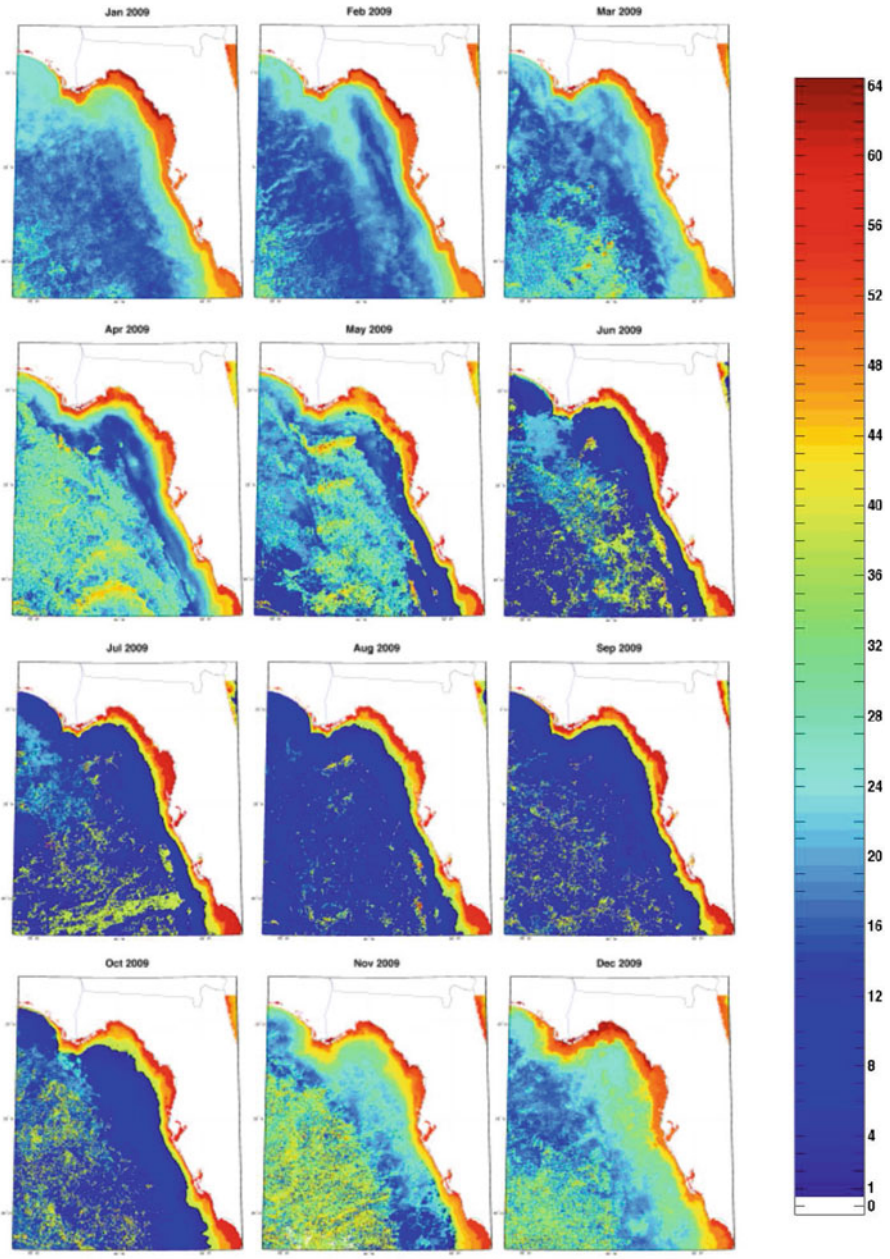

Fig. 21 A self-organizing map classification of Eastern Gulf of Mexico water for year 2009. We trained, and applied 5 years of monthly data from 2005 to 2009 using unsupervised artificial neural network technique called SOM. The color bar on the right shows the 64 classes corresponding to the colored region on the maps

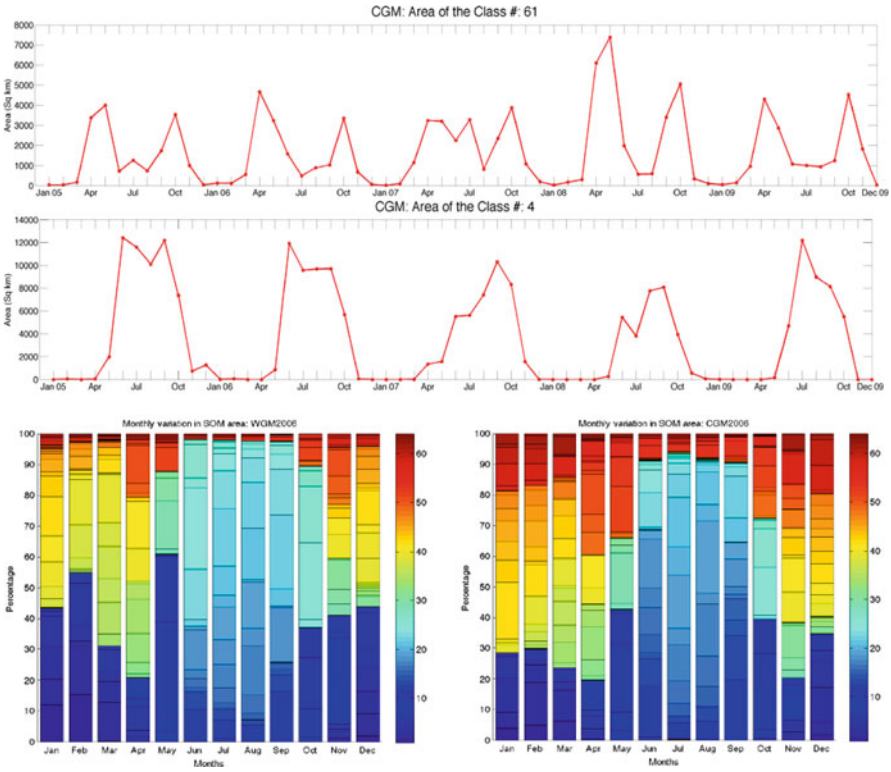

Fig. 22 The SOM classes show seasonal as well as annual trends. The area of SOM classes change depending upon the time of the year. For example, in the upper panels two classes are presented: Class 61 lies in the near-shore region and has a clear bi-annual cycle. The areal extent of this class is greatest in April and October, and least in January. Similarly, SOM class 4, which lies in the offshore region, shows a clear annual cycle. The area of class 4 has a maximum during the summer months and has a minimum during the winter months. The lower panel shows the monthly variation in the SOM class for 2006 for Western Gulf of Mexico (left) and Central Gulf of Mexico (right). The SOM classes in the range of 15–30 are dominant for the summer period (May–Sep)

Fish Catch and SOM Classes

The classes within the Self-organized Maps appear to correlate with fish counts. Figure 23 shows SEAMAP[8] fish count data for the western and central portions of the Gulf of Mexico in June 2006. The fish data are clearly clustered within classes 12–26. Those same classes dominate the SOM for June 2006. Further examination indicates that sea surface temperature is the driving factor for the dominant classes, with euphotic depth appearing to be the secondary driver.

[8] http://seamap.gsmfc.org/.

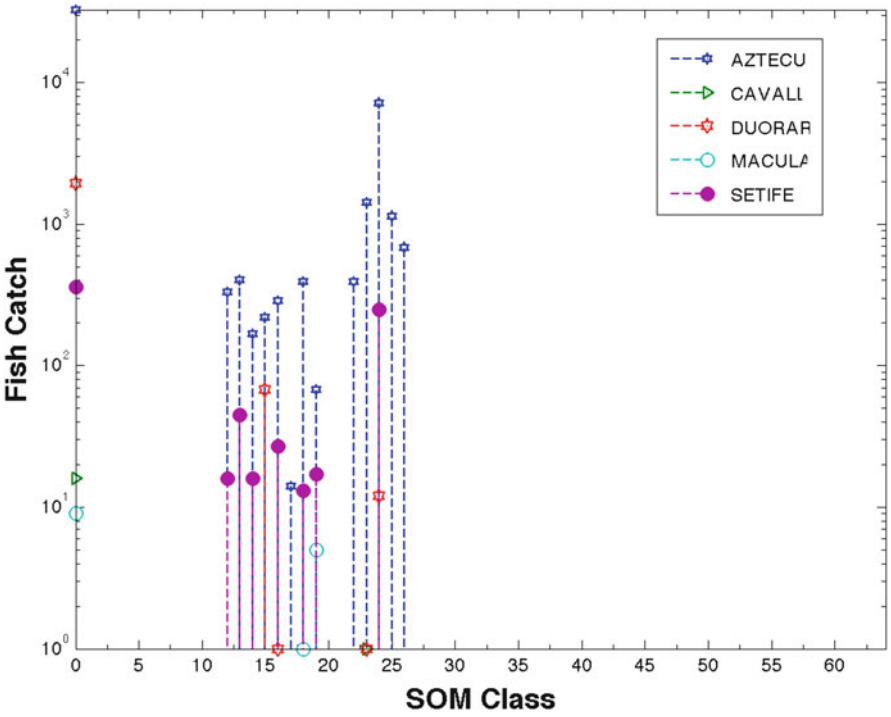

Fig. 23 Fish catch data from the Western Gulf of Mexico and Central Gulf of Mexico for June 2006 have been mapped to the SOM classes. The SOM classes in the range of 12–26 are the abundant classes and so is the fish catch

Some Likely Future Machine Learning Applications

Two recent advances are likely to open up a large number of new applications. First the improvement, size reduction and cost reduction of hyper-spectral imagery, and secondly, small embedded (credit card sized) GPU systems such as the NVIDIA Jetson TX1 with its 256 GPU cores.

Hyper-Spectral Imaging and Machine Learning for Real Time Embedded Processing and Decision Support

So what is Hyperspectral Imaging? The human eye perceives the color of visible light in three bands using the cones, the photoreceptor cells in the retina (Fig. 24). These three bands are red (centered on 564 nm), green (centered on 534 nm), and blue (centered on 420 nm). By contrast, instead of using just three broad bands, hyperspectral cameras divide the spectrum into a very large number of narrow

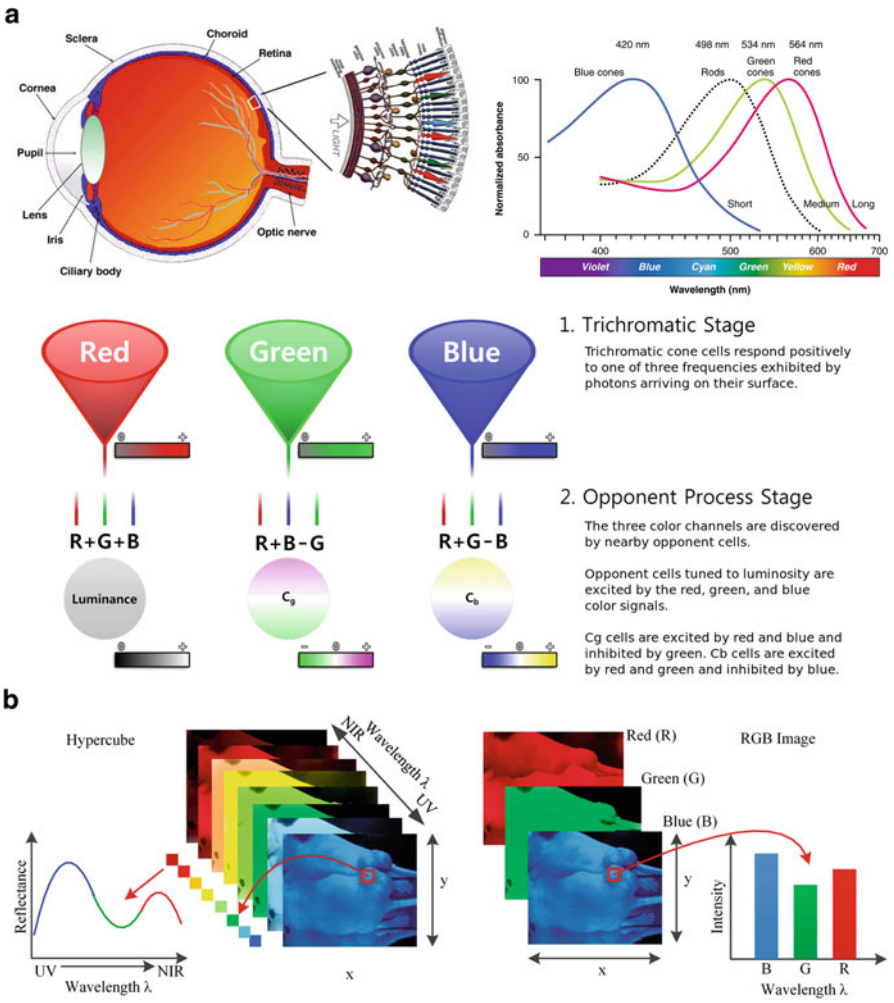

Fig. 24 Panel (**a**) Trichromatic cone cells in the eye respond to one of three wavelength ranges (RGB). These three bands are red (centered on 564 nm), green (centered on 534 nm), and blue (centered on 420 nm). Panel (**b**) shows a comparison between a hyperspectral-cube and RGB images. A hyper-cube is a three-dimensional dataset consisting of a two-dimensional image layers each for a different wavelength. So for each pixel in the image we have a multi-wavelength spectra (spectral signature). This is shown schematically in the lower left. On the right we see a conventional RGB color images with only three bands, images for red, green and blue wavelengths. The lower right shows an example 3 wavelength broad band spectra from a conventional RGB color image

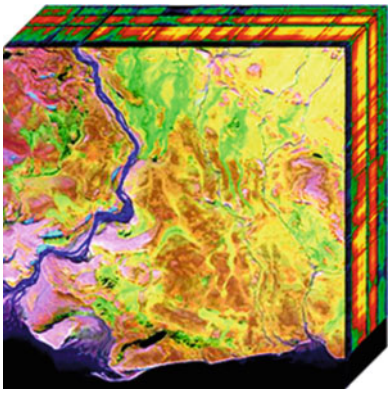

Fig. 25 Hyperspectral cube

bands. Sometimes as many as two to four hundred bands are used to create a hyperspectral datacube (Fig. 25). This technique of dividing images into bands can extend beyond the visible, into both the infrared and thermal infrared, and into the ultraviolet.

Hyperspectral imaging systems are used around the world in a variety of medical, laboratory, industrial, agricultural, and airborne applications. To illustrate the broader significance, let us briefly review just some of these (Fig. 27). Hyperspectral imaging (HSI) is used in various medical applications, especially in disease diagnosis and image-guided surgery. The disease diagnosis applications (e.g., skin examination) naturally lend themselves to telemedicine applications for rural communities where the network connectivity can drastically improve rural community medical care. For each snapshot in time, HSI acquires a three-dimensional dataset called a hypercube (Fig. 25), with two spatial dimensions (just like a regular camera) and one spectral dimension, there is a separate collocated image/layer for each wavelength band (Fig. 24).

Spatially resolved spectral imaging obtained by HSI can provide diagnostic information about the tissue physiology, morphology, and composition. With the advantage of acquiring two-dimensional images across a wide range of electromagnetic spectrum, HSI has been applied to numerous areas, including archaeology and art conservation (Angeletti et al., 2005; Liang, 2012), vegetation and water resource control (Govender et al., 2007), food quality and safety control (Gowen et al., 2007; Feng and Sun, 2012), forensic medicine (Malkoff and Oliver, 2000; Edelman et al., 2012), crime scene detection (Muller et al., 2003), biomedicine (Afromowitz et al., 1988; Carrasco et al., 2003), agriculture, security and defense, thin films, etc.

Figure 27 shows some of the many HSI applications. For example: Using an airborne HSI, an invasive weed ("leafy spurge", Euphorbia esula) infestation could be clearly identified (Jay et al., 2010), and a weed coverage map generated (Fig. 27a). A study of seed germination (Nansen et al., 2015) using HSI showed that although viable and nonviable seeds appear identical to the human eye they can be clearly distinguished using full reflectance spectra (Fig. 27b). Analysis of wound healing (La Fontaine et al., 2014) (Fig. 27c). Mapping hydrological

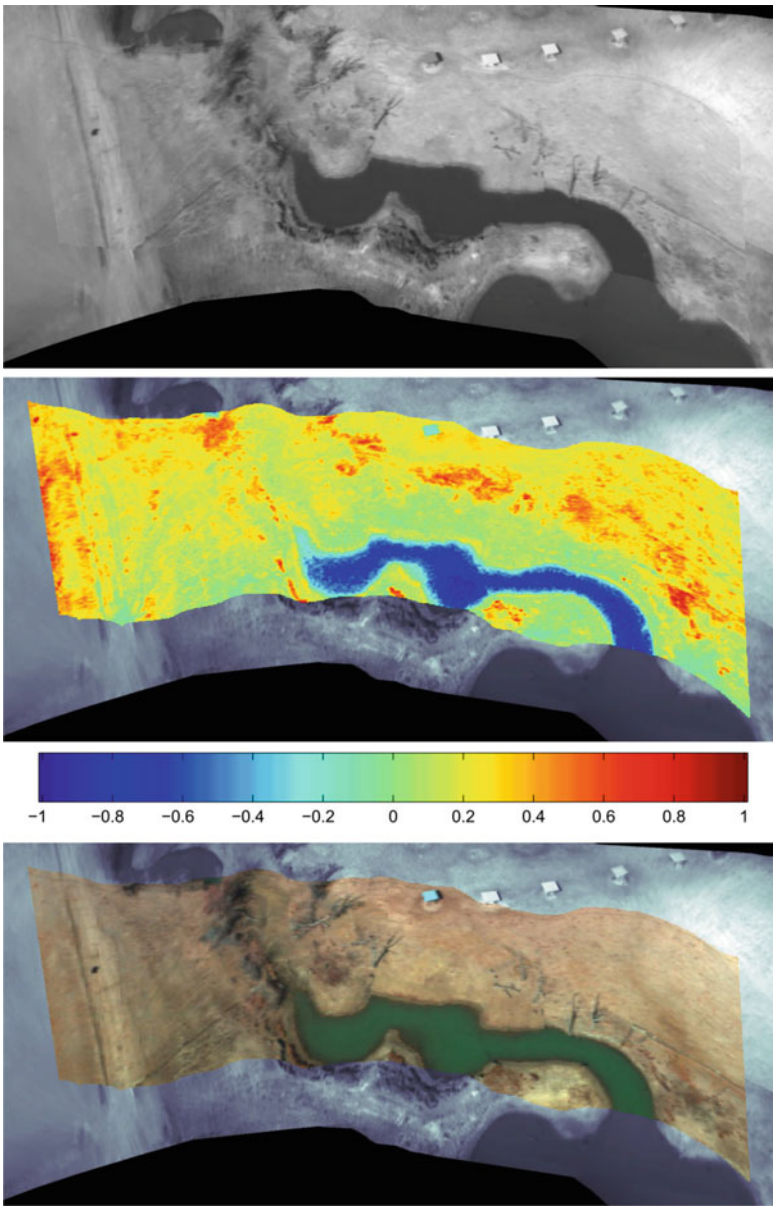

Fig. 26 Hyperspectral imaging of a rural landscape. Top image: sum of every spectral channel from the HS image, overlaid on top of the visible camera mosaic. Middle image: Normalized Difference Vegetation Index. Bottom image: pseudocolor from red, green, and blue channels (Ramirez, 2015)

formations (Fig. 27d). Fluorescent dye imaging (Fig. 27e). Examining the effect of surface pollution (Keith et al., 2009; Spangler et al., 2010) from leaking pipelines on vegetation (Fig. 27f). Checking food quality and fruit bruising (Fig. 27g). Classification of walnuts and shells (Fig. 27h). Automated analysis of cooked meats (Fig. 27i).

This diversity of examples demonstrates the general usefulness and applicability of HSI in a very broad range of contexts. In research, health, agriculture, industry, and more. We already saw in Sect. that combining the spectral signature in just seven wavelengths with machine learning was invaluable in uniquely identifying global dust sources with remarkable accuracy. So it can readily be seen that using more detailed hyper-spectral signatures with on-board embedded processing can provide incredibly powerful insights in a very compact package. Figure 26 shows an example of some hyperspectral imagery we obtained using our aerial vehicles (Ramirez, 2015). This approach is useful for many applications in smart agriculture, land surface classification, petrochemical surveying, disaster response (such as oil spills), etc. Let us take a closer look at the example of oil spill response.

Oil Spills

The National Academy of Sciences estimates 1.7–8.8 million tons of oil are released into global waters every year. More than 70% of this release is related to human activities. The effects of these spills include dead wildlife, oil covered marshlands and contaminated water (Fingas and Brown, 1997; Fingas, 2010; Liu et al., 2013; Cornwall, 2015). Spills of national significance (SONS), such as Deepwater Horizon (DWH), challenge response capabilities. In such large spills, *optimizing a coordinated response is a challenge*. There are always competing mission needs for aerial response resources such as helicopters and observer aircraft. Wildlife reconnaissance, oil observation overflights and targeting chemical dispersant application are a few examples. If we consider just one aspect, i.e. the spill itself, the challenges include both characterizing the continual temporal and spatial evolution of the spill extent, and the evolution of the oil itself as it weathers and emulsifies. Characterizing the oil spill can be made even more challenging due to the variable spill illumination and the weather. *Trained* observers are required, and their deployment needs can include a wide area, which is also challenging. Further, what is the optimal flight path(s) that should be used by the observers on each deployment to best meet the current needs and *anticipate the future evolution of the oil spill* to put in place any required preemptive measures or contingencies, such as shoreline pre-cleaning or protective boom deployment? During the DWH oil spill operational trajectory forecasting, maps of key areas for aerial observations to improve trajectory modeling were produced daily by NOAA for the overflight teams (Fig. 27).

The DWH oil spill and the associated impact monitoring was aided by extensive airborne and space-borne passive and active remote sensing (Fingas and Brown,

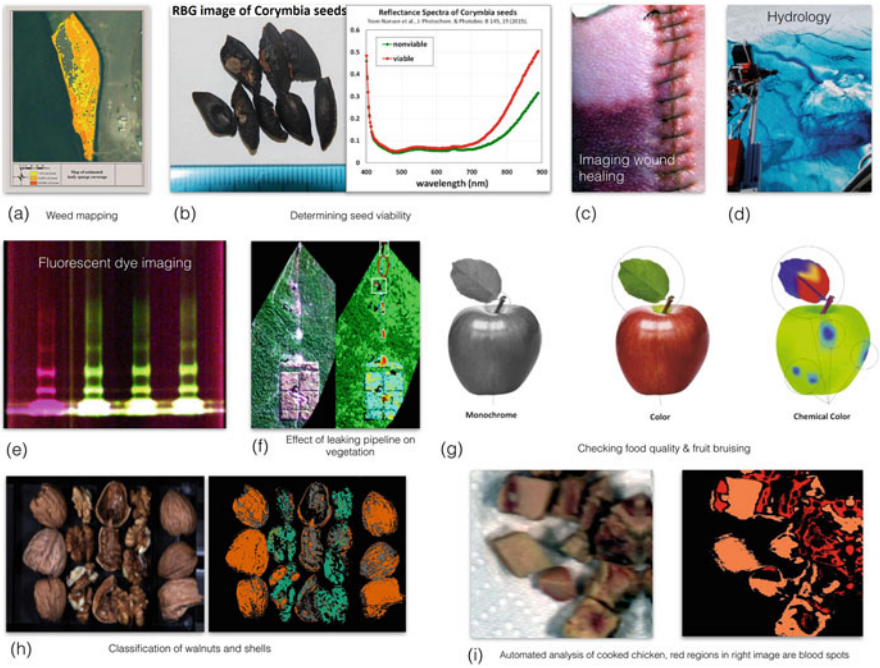

Fig. 27 Some examples of hyperspectral imaging applications

1997; Leifer et al., 2012; Liu et al., 2013; Fingas and Brown, 2014). A good review of these remote sensing activities is provided by (Leifer et al., 2012). During DWH, remote sensing was used to derive oil thickness (see Fig. 28) quantitatively for thick (>0.1 mm) slicks from AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) that measured 224 contiguous spectral bands with wavelengths from 400 to 2,500 nanometers (nm) using a spectral library approach based on the shape and depth of near infrared spectral absorption features (Kokaly et al., 2013; Leifer et al., 2012). MODIS (Moderate Resolution Imaging Spectroradiometer) satellite, visible-spectrum broadband data of surface-slick modulation of sunglint reflection allowed extrapolation to the total slick. A multispectral expert system used a neural network approach to provide Rapid Response thickness class maps (Sveykovsky and Muskat, 2006; Svejkovsky et al., 2009).

Airborne and satellite synthetic aperture radar (SAR) provides synoptic data under all-sky conditions (Liu et al., 2011; Leifer et al., 2012); however, SAR generally cannot discriminate thick (>100 μm) oil slicks from thin sheens (to 0.1 μm). The UAVSAR's (Unmanned Aerial Vehicle SAR) significantly greater signal-to-noise ratio and finer spatial resolution allowed successful pattern discrimination related to a combination of oil slick thickness, fractional surface coverage, and emulsification.

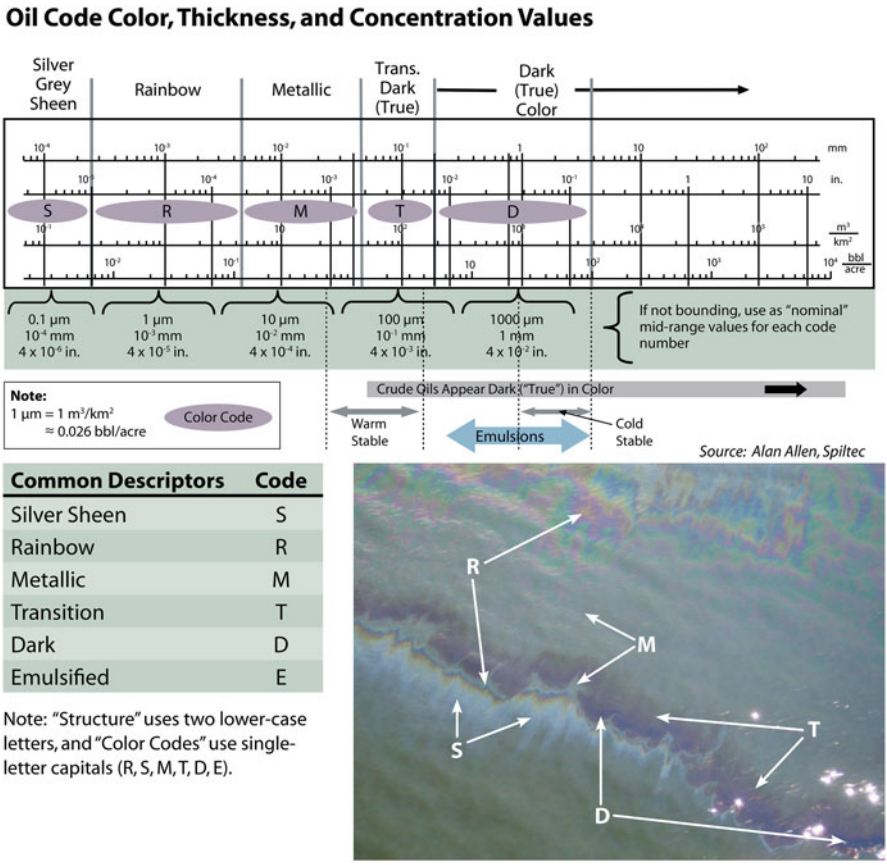

Fig. 28 Oil thickness chart and appearance from NOAA open water oil identification job aid

Further, in situ burning and smoke plumes were studied with AVIRIS and corroborated spaceborne CALIPSO (Cloud Aerosol Lidar and Infrared Pathfinder Satellite Observation) observations of combustion aerosols. CALIPSO and bathymetry lidar

data documented shallow subsurface oil, although ancillary data were required for confirmation.

Airborne hyperspectral, thermal infrared data have nighttime and overcast collection advantages, and were collected as well as MODIS thermal data. However, interpretation challenges and a lack of Rapid Response Products prevented significant use. Rapid Response Products were key to response utilization—*data needs are time critical*; thus, a high technological readiness level is vital to operational use of remote sensing products. The DWH oil spill experience demonstrated that development and operationalization of new near real-time spill response remote sensing tools must precede the next major oil spill (Leifer et al., 2012).

Cleanup of a SONS involve *multiple* skimmer ships, vessels collecting oil for in situ burning, and chemical dispersant operations. Typically these **slow** moving response ships are spread over a **large** area and guided by air support (e.g., helicopter), as the vessel bridge is too low to see the variation in thickness of the oil. Typically the manned air support will inform each ship of the location of recoverable oil ahead and then leave to overfly the next ship.

The cost of manned air support is significant, so each ship does *not* usually have dedicated continuous manned air support. In smaller spills, a report of oil location is given to the responding ship, these ships move slowly, so by the time they reach the location provided by manned air support, the oil has *moved*! For oil cleanup to be optimal (quickest) and effective (most oil recovered) the skimmer ships need to first focus on the regions of thickest oil. From the visual perspective of the ships crew, relatively close to the water and with a shallow viewing angle, it is not easy to know where the thickest oil is. Accurately discerning the gradations in black oil thickness is a challenging task.

Rich information on the thickness of the oil and the degree of weathering is contained in the detailed hyperspectral signature of the oil spill. This information can be utilized through the use of machine learning to provide real time response tools.

Summary

We have seen that machine learning has found many applications in remote sensing. These applications range from retrieval algorithms to bias correction, from code acceleration to detection of disease in crops, from classification of pelagic habitats to the rock type classification. As a broad subfield of artificial intelligence, machine learning is concerned with algorithms and techniques that allow computers to "learn." The major focus of machine learning is to extract information from data automatically by computational and statistical methods. Over the last decade there has been considerable progress in developing a machine learning methodology for a variety of Earth Science applications involving trace gases, retrievals, aerosol products, land surface products, vegetation indices, and most recently, ocean applications.

References

7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/. Accessed 29 Aug 2016

Afromowitz MA, Callis JB, Heimbach DM, Desoto LA, Norton MK (1988) Multispectral imaging of burn wounds - a new clinical instrument for evaluating burn depth. IEEE Trans Biomed Eng 35(10):842–850. https://doi.org/10.1109/10.7291. <GotoISI>://WOS:A1988Q389400009

Andrews CP, Ratner PH, Ehler BR, Brooks EG, Pollock BH, Ramirez DA, Jacobs RL (2013). The mountain cedar model in clinical trials of seasonal allergic rhinoconjunctivitis. Ann Allergy Asthma Immunol 111(1):9–13

Angeletti C, Harvey NR, Khomitch V, Fischer AH, Levenson RM, Rimm DL: Detection of malignancy in cytology specimens using spectral-spatial analysis. Lab Investig 85(12):1555–1564. https://doi.org/10.1038/labinvest.3700357. <GotoISI>://WOS:000233372200011

Arizmendi C, Sanchez J, Ramos N, Ramos G (1993) Time series predictions with neural nets: application to airborne pollen forecasting. Int J Biometeorol 37(3):139–144

Aurin M, Mannino A (2012) A database for developing global ocean color algorithms for colored dissolved organic material, CDOM spectral slope, and dissolved organic carbon, paper presented at Ocean Optics XXI, The Oceanography Society, Glasgow, UK

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford (1995)

Bouet C, Cautenet G, Bergametti G, Marticorena B, Todd MC, Washington R (2005) Sensitivity of desert dust emissions to model horizontal grid spacing during the Bodele Dust Experiment 2005. Atmos Environ 50, 377–380 (2012)

Breiman L (1984) Classification and regression trees. The Wadsworth statistics/probability series. Wadsworth International Group, Belmont, CA

Breiman L (2001) Random forests. Mach Learn 45(1):5–32 (2001). Times Cited: 3621

Britton J, Pavord I, Richards K, Knox A, Wisniewski A, Wahedna I, Kinnear W, Tattersfield A, Weiss S (1994) Factors influencing the occurrence of airway hyperreactivity in the general population: the importance of atopy and airway calibre. Eur Respir J 7(5):881–887

Brown ME, Lary DJ, Vrieling A, Stathakis D, Mussa H (2008) Neural networks as a tool for constructing continuous ndvi time series from AVHRR and MODIS. Int J Remote Sens 29(24):7141–7158

Carrasco O, Gomez R, Chainani A, Roper W (2003) Hyperspectral imaging applied to medical diagnoses and food safety. In: Proceedings of the society of photo-optical instrumentation engineers (SPIE), vol 5097, pp. 215–221. https://doi.org/10.1117/12.502589 <GotoISI>://WOS:000185395500023

Castellano-Méndez M, Aira M, Iglesias I, Jato V, González-Manteiga W (2005) Artificial neural networks as a useful tool to predict the risk level of betula pollen in the air. Int J Biometeorol 49(5):310–316

Committee FGD, et al. (2012) Coastal and marine ecological classification standard. Publication# FGDC-STD-018-2012

Cornwall W (2015) Deepwater horizon: after the oil. Science 348(6230):22–29. https://doi.org/10.1126/science.348.6230.22. http://www.sciencemag.org/content/348/6230/22.short

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. Times Cited: 3429

Csépe Z, Makra L, Voukantsis D, Matyasovszky I, Tusnády G, Karatzas K, Thibaudon M (2014) Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. Sci Total Environ 476:542–552

D'amato G, Spieksma FTM (1991) Allergenic pollen in Europe. Grana 30(1):67–70

Demuth HB, Beale MH, De Jess O, Hagan MT (2014) Neural network design, 2nd edn. Martin Hagan, Stillwater (2014)

Domingos P (2015) The master algorithm: how the quest for the ultimate learning machine will remake our world. Basic Books, New York

Edelman GJ, Gaston E, van Leeuwen TG, Cullen PJ, Aalders MCG (2012) Hyperspectral imaging for non-contact analysis of forensic traces. Forensic Sci Int 223(1–3):28–39. https://doi.org/10.1016/j.forsciint.2012.09.012. <GotoISI>://WOS:000311432100021

Ernst P, Ghezzo H, Becklake M (2002) Risk factors for bronchial hyperresponsiveness in late childhood and early adolescence. Eur Respir J 20(3):635–639

Esch RE, Hartsell CJ, Crenshaw R, Jacobson RS (2001) Common allergenic pollens, fungi, animals, and arthropods. Clin Rev Allergy Immunol 21(2):261–292

Feng YZ, Sun DW (2012) Application of hyperspectral imaging in food safety inspection and control: a review. Crit Rev Food Sci Nutr 52(11):1039–1058. https://doi.org/10.1080/10408398.2011.651542. <GotoISI>://WOS:000306740000007

Fingas M (2010) Oil spill science and technology. Gulf Professional Publishing, Houston (2010)

Fingas MF, Brown CE (1997) Review of oil spill remote sensing. Spill Sci Technol Bull 4(4):199–208. http://dx.doi.org/10.1016/S1353-2561(98)00023-1. http://www.sciencedirect.com/science/article/pii/S1353256198000231. The Second International Symposium on Oil Spills

Fingas M, Brown C (2014) Review of oil spill remote sensing. Mar Pollut Bull 83(1):9–23. http://dx.doi.org/10.1016/j.marpolbul.2014.03.059. http://www.sciencedirect.com/science/article/pii/S0025326X14002021

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning. Springer Series in Statistics, vol 1. Springer, Berlin

Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recogn Lett 31(14):2225–2236

Govender M, Chetty K, Bulcock H (2007) A review of hyperspectral remote sensing and its application in vegetation and water resource studies. Water SA 33(2):145–151. <GotoISI>://WOS:000246960100001

Gowen AA, O'Donnell CP, Cullen PJ, Downey G, Frias JM (2007) Hyperspectral imaging - an emerging process analytical tool for food quality and safety control. Trends Food Sci Technol 18(12):590–598. https://doi.org/10.1016/j.tifs.2007.06.001. <GotoISI>://WOS:000251485800001

Haykin SS (1994) Neural networks: a comprehensive foundation. Macmillan, New York. 93028092 Simon Haykin. ill. ; 26 cm. Includes bibliographical references (pp 635–690) and index

Haykin SS (1999) Neural networks: a comprehensive foundation, 2nd edn. Prentice Hall, Upper Saddle River, NJ. 98007011 Simon Haykin. ill. ; 25 cm. Includes bibliographical references (pp 796–836) and index

Haykin SS (2001) Kalman filtering and neural networks. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York. 2001049240 edited by Simon Haykin. ill. ; 24 cm. "A Wiley Interscience publication." Includes bibliographical references and index

Haykin SS (2007) New directions in statistical signal processing : from systems to brain. Neural information processing series. MIT Press, Cambridge, MA. 2005056210 GBA671791 013536699 (OCoLC)ocm62302576 (OCoLC)62302576 edited by Simon Haykin . . . [et al.]. ill. ; 26 cm. Includes bibliographical references (p. [465]–508) and index. Modeling the mind : from circuits to systems/Suzanna Becker – Empirical statistics and stochastic models for visual signals/David Mumford – The machine cocktail party problem/Simon Haykin, Zhe Chen – Sensor adaptive signal processing of biological nanotubes (ion channels) at macroscopic and nano scales/Vikram Krishnamurthy – Spin diffusion : a new perspective in magnetic resonance imaging/Timothy R. Field – What makes a dynamical system computationally powerful?/Robert Legenstein, Wolfgang Maass – A variational principle for graphical models/Martin J. Wainwright, Michael I. Jordan – Modeling large dynamical systems with dynamical consistent neural networks/Hans-Georg Zimmermann . . . [et al.] – Diversity in communication : from source coding to wireless networks/Suhas N. Diggavi – Designing patterns for easy recognition : information transmission with low-density parity-check codes/Frank R. Kschischang, Masoud Ardakani – Turbo processing/Claude Berrou, Charlotte Langlais, Fabrice Seguin – Blind signal processing based on data geometric properties/Konstantinos Diamantaras – Game-theoretic learning/Geoffrey J. Gordon – Learning observable operator models via the efficient sharpening algorithm/Herbert Jaeger . . . [et al.]

Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844

Howard LE, Levetin E (2014) Ambrosia pollen in tulsa, oklahoma: aerobiology, trends, and forecasting model development. Ann Allergy Asthma Immunol 113(6):641–646

Jay SC, Lawrence RL, Repasky KS, Rew LJ (2010) IEEE: detection of leafy spurge using hyperspectral-spatial-temporal imagery. In: 2010 IEEE international geoscience and remote sensing symposium, pp 4374–4376. https://doi.org/10.1109/igarss.2010.5652580. <GotoISI>://WOS:000287933804134

Kasprzyk I (2008) Non-native ambrosia pollen in the atmosphere of rzeszów (se poland); evaluation of the effect of weather conditions on daily concentrations and starting dates of the pollen season. Int J Biometeorol 52(5):341–351

Keith CJ, Repasky KS, Lawrence RL, Jay SC, Carlsten JL (2009) Monitoring effects of a controlled subsurface carbon dioxide release on vegetation using a hyperspectral imager. Int J Greenhouse Gas Control 3(5):626–632. https://doi.org/10.1016/j.ijggc.2009.03.003. <GotoISI>://WOS:000268949300013

Kinney PL (2008) Climate change, air quality, and human health. Am J Prev Med 35(5):459–467

Kokaly RF, Couvillion BR, Holloway JM, Roberts DA, Ustin SL, Peterson SH, Khanna S, Piazza SC (2013) Spectroscopic remote sensing of the distribution and persistence of oil from the deepwater horizon spill in barataria bay marshes. Remote Sens Environ 129:210–230. http://dx.doi.org/10.1016/j.rse.2012.10.028. http://www.sciencedirect.com/science/article/pii/S0034425712004166

Koren I, Kaufman YJ, Washington R, Todd MC, Rudich Y, Martins JV, Rosenfeld D (2006) The Bodele depression: a single spot in the Sahara that provides most of the mineral dust to the Amazon forest. Environ Res Lett 1(1):011001

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: A review of classification techniques

Laaidi M, Laaidi K, Besancenot JP, Thibaudon M (2003) Ragweed in france: an invasive plant and its allergenic pollen. Ann Allergy Asthma Immunol 91(2):195–201

La Fontaine J, Lavery L, Zuzak K (2014) The use of hyperspectral imaging (HSI) in wound healing. Proc SPIE 8979. https://doi.org/897903 10.1117/12.2041841. <GotoISI>://WOS:000336032300001

Lary D, Aulov, O (2008) Space-based measurements of HCl: intercomparison and historical context. J Geophys Res: Atmos 113(D15)

Lary D, Müller M, Mussa H (2003) Using neural networks to describe tracer correlations. Atmos Chem Phys Discuss 3(6):5711–5724

Lary D, Waugh D, Douglass A, Stolarski R, Newman P, Mussa H (2007) Variations in stratospheric inorganic chlorine between 1991 and 2006. Geophys Res Lett 34(21)

Lary DJ, Remer LA, MacNeill D, Roscoe B, Paradise S (2009) Machine learning and bias correction of MODIS aerosol optical depth. IEEE Geosci Remote Sens Lett 6(4):694–698

Lary DJ, Faruque FS, Malakar N, Moore A, Roscoe B, Adams ZL, Eggelston Y (2014) Estimating the global abundance of ground level presence of particulate matter (pm2. 5). Geospat Health 8(3):611–630

Lary DJ, Alavi AH, Gandomi AH, Walker AL (2016) Machine learning in geosciences and remote sensing. Geosci Front 7(1):3–10

Leifer I, Lehr WJ, Simecek-Beatty D, Bradley E, Clark R, Dennison P, Hu Y, Matheson S, Jones CE, Holt B, Reif M, Roberts DA, Svejkovsky J, Swayze G, Wozencraft J (2012) State of the art satellite and airborne marine oil spill remote sensing: application to the BP deepwater horizon oil spill. Remote Sens Environ **124**:185–209. http://dx.doi.org/10.1016/j.rse.2012.03.024. http://www.sciencedirect.com/science/article/pii/S0034425712001563

Lellis-Dibble KA, McGlynn K, Bigford TE (2008) Estuarine fish and shellfish species in US commercial and recreational fisheries: economic value as an incentive to protect and restore estuarine habitat. National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Office of Habitat Conservation, Habitat Protection Division (2008)

Lewis WH, Vinay P, Zenger VE (1983) Airborne and allergenic pollen of North America. Johns Hopkins University Press, Baltimore

Liang H (2012) Advances in multispectral and hyperspectral imaging for archaeology and art conservation. Appl Phys A-Mater Sci Process 106(2):309–323

Liu P, Li X, Qu JJ, Wang W, Zhao C, Pichel W (2011) Oil spill detection with fully polarimetric {UAVSAR} data. Mar Pollut Bull 62(12):2611–2618. http://dx.doi.org/10.1016/j.marpolbul.2011.09.036. http://www.sciencedirect.com/science/article/pii/S0025326X11005248

Liu Y, MacFadyen A, Ji Z, Weisberg R (2013) Monitoring and modeling the deepwater horizon oil spill: a record breaking enterprise. Geophysical monograph series. Wiley, Hoboken

Low RB, Bielory L, Qureshi AI, Dunn V, Stuhlmiller DF, Dickey DA (2006) The relation of stroke admissions to recent weather, airborne allergens, air pollution, seasons, upper respiratory infections, and asthma incidence, september 11, 2001, and day of the week. Stroke 37(4):951–957

Malkoff DB, Oliver WR (2000) Hyperspectral imaging applied to forensic medicine. Progr Biomed Opt 1:108–116. <GotoISI>://WOS:000086469300013

Matheson EM, Player MS, Mainous AG, King DE, Everett CJ (2008) The association between hay fever and stroke in a cohort of middle aged and elderly adults. J Am Board Fam Med 21(3):179–183

McCulloch W, Pitts W (1943) Bull Math Biophys 5:115. https://doi.org/10.1007/BF02478259

Muller MG, Valdez TA, Georgakoudi I, Backman V, Fuentes C, Kabani S, Laver N, Wang ZM, Boone CW, Dasari RR, Shapshay SM, Feld MS (2003) Spectroscopic detection and evaluation of morphologic and biochemical changes in early human oral carcinoma. Cancer 97(7):1681–1692. https://doi.org/10.1002/cncr.11255. <GotoISI>://WOS:000181816600012

Nansen C, Zhao G, Dakin N, Zhao C, Turner SR (2015) Using hyperspectral imaging to determine germination of native Australian plant seeds. J Photochem Photobiol B-Biol 145:19–24. https://doi.org/10.1016/j.jphotobiol.2015.02.015. <GotoISI>://WOS:000352679200003

Nowosad J (2016) Spatiotemporal models for predicting high pollen concentration level of corylus. Int J Biometeorol. Springer 60(6):843–855

Osowski S, Garanty K (2007) Forecasting of the daily meteorological pollution using wavelets and support vector machine. Eng Appl Artif Intell 20(6):745–755

Oswalt ML, Marshall GD (2008) Ragweed as an example of worldwide allergen expansion. Allergy Asthma Clin Immunol 4(3):130

Puc M (2012) Artificial neural network model of the relationship between betula pollen and meteorological factors in szczecin (Poland). Int J Biometeorol 56(2):395–401

Ramirez DA (1984) The natural history of mountain cedar pollinosis. J Allergy Clin Immunol 73(1):88–93

Ramirez JP, Lary DJ, Gans NR (2015) Low-altitude terrestrial spectroscopy from a pushbroom sensor. J Field Robot 1–16. https://doi.org/10.1002/rob.21624

Rodríguez-Rajo F, Astray G, Ferreiro-Lage J, Aira M, Jato-Rodriguez M, Mejuto JC (2010) Evaluation of atmospheric poaceae pollen concentration using a neural network applied to a coastal atlantic climate region. Neural Netw 23(3):419–425

Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674. https://doi.org/10.1109/21.97458

Sánchez-Mesa J, Galán C, Martínez-Heras J, Hervás-Martínez C (2002) The use of a neural network to forecast daily grass pollen concentration in a mediterranean region: the southern part of the iberian peninsula. Clin Exp Allergy 32(11):1606–1612

Sassen K (2008) Boreal tree pollen sensed by polarization lidar: depolarizing biogenic chaff. Geophys Res Lett 35(18):L18810

Spangler LH, Dobeck LM, Repasky KS, Nehrir AR, Humphries SD, Barr JL, Keith CJ, Shaw JA, Rouse JH, Cunningham AB, Benson SM, Oldenburg CM, Lewicki JL, Wells AW, Diehl JR, Strazisar BR, Fessenden JE, Rahn TA, Amonette JE, Barr JL, Pickles WL, Jacobson JD, Silver EA, Male EJ, Rauch HW, Gullickson KS, Trautz R, Kharaka Y, Birkholzer J, Wielopolski L (2010) A shallow subsurface controlled release facility in bozeman, montana, usa, for testing near surface CO2 detection techniques and transport models. Environ Earth Sci 60(2):227–239. https://doi.org/10.1007/s12665-009-0400-2. <GotoISI>://WOS:000276637000002

Stark PC, Ryan LM, McDonald JL, Burge HA (1997) Using meteorologic data to predict daily ragweed pollen levels. Aerobiologia 13(3):177–184
Svejkovsky J, Muskat J, Mullin J (2009) Adding a multispectral aerial system to the oil spill response arsenal. Sea Technol 50(8):17–22
Sveykovsky J, Muskat S (2006) Real-time detection of oil slick thickness patterns with a portable multispectral sensor. Tech. rep. (2006)
Todd MC, Washington R, Martins JV, Dubovik O, Lizcano G, M'Bainayel S, Engelstaedter S (2007) Mineral dust emission from the bodele depression, northern chad, during bodex 2005. J Geophys Res-Atmos 112(D6):D06207
Tränkle E, Mielke B (1994) Simulation and analysis of pollen coronas. Appl Opt 33(21): 4552–4562
Vapnik VN (1982) Estimation of dependences based on empirical data. Springer series in statistics. Springer, New York
Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
Vapnik VN (2000) The nature of statistical learning theory. Statistics for engineering and information science, 2nd edn. Springer, New York
Vapnik VN (2006) Estimation of dependences based on empirical data ; empirical inference science : afterword of 2006. Information science and statistics, 2nd enl. edn. Springer, New York, NY
Voukantsis D, Niska H, Karatzas K, Riga M, Damialis A, Vokou D (2010) Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. Atmos Environ 44(39):5101–5111
Washington R, Todd MC (2005) Atmospheric controls on mineral dust emission from the bodele depression, chad: the role of the low level jet. Geophys Res Lett 32(17):4543
Washington R, Todd MC, Engelstaedter S, Mbainayel S, Mitchell F (2006) Dust and the low-level circulation over the bodele depression, chad: observations from bodex 2005. J Geophys Res-Atmos **111**(D3):D03201
Washington R, Todd MC, Lizcano G, Tegen I, Flamant C, Koren I, Ginoux P, Engelstaedter S, Bristow CS, Zender CS, Goudie AS, Warren A, Prospero JM (2006) Links between topography, wind, deflation, lakes and dust: the case of the Bodele depression, chad. Geophys Res Lett 33(9):L09401
Wayne P, Foster S, Connolly J, Bazzaz F, Epstein P (2002) Production of allergenic pollen by ragweed (ambrosia artemisiifolia l.) is increased in CO2-enriched atmospheres. Ann Allergy Asthma Immunol 88(3):279–282
Zhao F, Elkelish A, Durner J, Lindermayr C, Winkler JB, Ruëff, F., Behrendt, H., Traidl-Hoffmann, C., Holzinger, A., Kofler, W., et al.: Common ragweed (*Ambrosia artemisiifolia* L.): allergenicity and molecular characterization of pollen after plant exposure to elevated NO2. Plant Cell Environ 39(1):147–164