Predicting the spectral information of future land cover using machine learning

Patil, Sopan; Gu, Yuting; Dias, A.; Steiglitz, Marc; Turk, Greg

International Journal of Remote Sensing

DOI:
10.1080/01431161.2017.1343512

Published: 01/01/2017

Peer reviewed version

Predicting the spectral information of future land cover using machine learning

Sopan D. Patil[1], Yuting Gu[2], Felipe S. A. Dias[3], Marc Stieglitz[3], Greg Turk[2]

[1] School of Environment, Natural Resources and Geography,
Bangor University,
Deiniol Road, Bangor, LL57 2UW, United Kingdom

[2] School of Interactive Computing,
Georgia Institute of Technology,
85 Fifth Street NW, Atlanta, GA 30308, United States of America

[3] School of Civil and Environmental Engineering,
Georgia Institute of Technology,
790 Atlantic Drive, Atlanta, GA 30332, United States of America

Submission to: International Journal of Remote Sensing

Corresponding author: Sopan D. Patil (email: s.d.patil@bangor.ac.uk, Tel: +44 1248388294)

Funding information: This work was supported by the National Science Foundation under Award Number 1027870 (CDI-Type Small Resources Supercomputing: High Performance Computing in the Earth Sciences).

Abstract

Application of machine learning models to study land cover change is typically restricted to the change detection of categorical, i.e., classified, land cover data. In this study, our aim is to extend the utility of such models to predict the spectral band information of satellite images. A Random Forests (RF) based machine learning model is trained using topographic and historical climatic variables as inputs to predict the spectral band values of high-resolution satellite imagery across two large sites in the western United States, New Mexico (10,570 km^2) and Washington (9,400 km^2). The model output is used to obtain a true colour photorealistic image and an image showing the Normalized Difference Vegetation Index (NDVI) values. We then use the trained model to explore what the land cover might look like for a climate change scenario during the 2061-2080 period. The RF model achieves high validation accuracy for both sites during the training phase, with the coefficient of determination (R^2) = 0.79 for New Mexico site and R^2 = 0.73 for Washington site. For the climate change scenario, prominent land cover changes are characterized by an increase in the vegetation cover at the New Mexico site and a decrease in the perennial snow cover at the Washington site. Our results suggest that direct prediction of spectral band information is highly beneficial due to the ability it provides for deriving ecologically relevant products, which can be used to analyse land cover change scenarios from multiple perspectives.

Keywords: Land cover change; climate change; machine learning; Random forest; Landsat.

1 Introduction

Recent warming of the climate has led to large-scale changes in earth's land cover. Large scale warming has resulted in a shift in the dominant vegetation species to higher latitudes and higher elevations, which has been reported in many parts of the world [*Walther et al.*, 2002; *Root et al.*, 2003; *Kelly and Goulden*, 2008; *Lenoir et al.*, 2008; *VanDerWal et al.*, 2013]. Throughout the southwest US, woody species have been encroaching on grasslands [*Barger et al.*, 2011]. In southwestern Wyoming, where precipitation has been trending down for the last thirty years, sagebrush vegetation have been giving way to bare ground [*Homer et al.*, 2015]. In many western states of the US, where seasonal snowmelt accounts for a large fraction of the annual water supply, winter snow accumulation and perennial snow cover has been decreasing. *Mote* [2003] has shown that from the mid to latter half of the twentieth century, winter snow accumulation at several locations along the Cascades Mountain Range fell by more than 40%. *Hall et al.* [2015] have reported that in north-western Wyoming the winter snowmelt is 16 ± 10 days earlier in 2000s compared to the period 1972 - 1999. At higher latitudes, where warming has been significantly greater than the planetary average, there has been simultaneous shortening of the snow season [*Groisman et al.*, 1994; *Stow et al.*, 2004] and lengthening of the vegetation growing season [*Foster*, 1989; *Foster et al.*, 1992; *Stone et al.*, 2002]. These are just some of the land cover changes that studies have documented within the last 100 years. However, while this evidence of change provides a view to the future change, it nevertheless remains highly uncertain what changes will occur in the global land cover over the next 100 years.

Despite high uncertainty, numerous studies have attempted to model the potential impact of climate change on future land cover [*Pearson and Dawson*, 2003; *Sitch et al.*, 2003, 2008; *Krinner et al.*, 2005; *Rogan et al.*, 2008]. We can broadly classify these modelling efforts into those using physically based and statistically based models. Physically based models provide a mechanistic framework in which mathematical representation of individual processes, such as vegetation growth and decline, snow dynamics, and land-atmosphere exchanges of water and carbon, can be coupled to simulate an integrated landscape response to climate forcing. For instance, *Sitch et al.* [2003] developed the Lund-Potsdam-Jena (LPJ) Dynamic Global Vegetation Model (DGVM) to simulate the response of terrestrial vegetation to climate forcing and demonstrated its application globally at $0.5° \times 0.5°$ spatial resolution. *Campbell et al.* [2010] used the Simultaneous Heat and Water (SHAW) model to simulate future changes in snowpack and soil frost at the Hubbard Brook Experimental Forest in New Hampshire, USA with climate forcing from three different General Circulation Models (GCMs). Physically based models have the benefit that they can be used to infer the cause and effect of land cover change at the level of individual physical processes [*Parker et al.*, 2003; *Pauleit et al.*, 2005; *Pitman et al.*, 2009]. However, these models suffer from the large number of simulations necessary to adequately constrain parameter values, and therefore can be both time consuming and, in many instances, beyond the available computing power for many researchers. As a result, physically based simulations tend to make a compromise in their spatial resolution [*Brovkin et al.*, 2006; *Verburg et al.*, 2011] or their areal extent [*Tague et al.*, 2009; *Abdelnour et al.*, 2011, 2013].

Statistically based land cover change models, on the other hand, operate on the premise that a strong relationship exists between the geographical distribution of land cover and the environmental and climate conditions and that these relationships can be empirically extracted using statistical machine learning methods [*DeFries and Chan*, 2000; *Guisan and Zimmermann*, 2000; *McIver and Friedl*, 2002; *Brown de Colstoun et al.*, 2003; *Guisan et al.*, 2006; *Klein et al.*, 2012]. Machine learning refers to a broad set of computational techniques used for identifying patterns in data and are usually applied where standard techniques such as regression analysis are not applicable. Machine learning algorithms statistically learn patterns and rules based on present correlations defined by a training set of data and provides a learned mapping between predictor variables (or attributes) and a target variable [*Witten et al.*, 1993; *Bishop*, 2006]. Once a model is developed through training, it can be used to predict the target variable in situations where the predictor variables are known but the target variable is not [*Mitchell*, 1997]. Some of the widely used machine learning techniques include Neural Networks (NN), Support Vector Machines (SVM), Classification Trees (CT), Regression Trees (RT), Random Forests (RF), Boosted Regression Trees (BRT), and Multivariate Adaptive Regression Splines [*Vapnik*, 1999; *Domingos*, 2012; *Alpaydin*, 2014].

Machine learning models have been widely used to predict the changes in land cover for a given site or region. *Rogan et al.* [2008] compared three different machine learning models (CT, Maximum Likelihood Classification, and NN) to detect changes in land cover classes across two sites in California, USA between the years 1990 and 1996. Similar model comparison was done by *Schneider* [2012] for land cover change detection in China across five time periods between 1988 and 2009. *Pearson et al.* [2013] used a RF model to identify relationships between 19 bioclimatic variables from the WorldClim database and eight tundra vegetation types in the Arctic, and then used the trained model to predict future vegetation cover classes for the climate change scenarios in the 2050s. Statistical machine learning models have an advantage over the physically based models due to their significantly faster computational speed and better predictive capacity [*Im and Jensen*, 2005; *Rogan et al.*, 2008]. Thus, they can operate at both high spatial resolutions and over very large areas with much lower computational overhead. However, one limitation of the machine learning models is that their application has so far been restricted to the change detection/prediction of categorical (i.e., classified) land cover information.

In this paper, our goal is to extend the utility of machine learning models to predict the spectral band information of high-resolution satellite based land cover images (which is continuous scale numerical data) for a future climate change scenario. The rationale for doing so is two-fold. First, there is a body of evidence that strongly relates remote sensing proxies, such as the Normalized Difference Vegetation Index (NDVI), to ecologically important processes [*Roughgarden et al.*, 1991; *Kerr and Ostrovsky*, 2003; *Pettorelli et al.*, 2005], and their prediction into the future will offer a quantitative understanding of ecological change. Availability of spectral band information for a future scenario would be critical to derive such proxy data. Second, as will be demonstrated, our methodology can be used to provide a photorealistic view of land cover change, which from a conceptual vantage point, provides new and intuitive insights to understand the implications of change. To conduct this research, we have used the topographic and historical climate data (1950-2000) from two large sites in the United States, one in the state of New Mexico and the other in the state of Washington, to train a RF machine learning model. The model is trained to predict the spectral values from bands 1 (Blue), 2 (Green), 3 (Red), and 4 (Near Infrared) of a Landsat 7 image. Then, with the GCM climate forecast data from the 2061-2080 period as input, we use the trained model to predict the future band information and its derivative RGB and NDVI images. The data used include Landsat 7 reflectance imagery, mean annual temperature and annual precipitation for the 1950-2000 period, Digital Elevation Model data, and the future climate projections generated using the Goddard Institute for Space Studies (GISS) GCM version E2 that are downscaled and bias corrected to the current climate.

2 Study Area and Data

2.1 Study sites

Our New Mexico site (Figure 1) is located in the north-central region of New Mexico state in the western US and includes the N-S flowing Rio Grande River, the Jimenez Mountains on the west, and the Santa Fe National Forest on the east. Elevation ranges from 1573 to 3972 m. The annual mean temperature ranges from -1.3°C at higher altitudes to 12.7°C in the valleys. Annual precipitation ranges from 250 mm in the valleys to 1000 mm in the uplands. Dominant vegetation types include grasses near the river channel, shrubs in the lowlands and along the mountain slopes, and evergreen vegetation in the uplands. Uplands also include grasses and a small fraction of mixed forest. The soil type in this region consists mainly of Entisols, Inceptisols and Alisols. Exposed rock formations are also present in areas surrounding the mountain peaks [*Wolock*, 1997]. The total area covered is 10,570 km^2.

[Insert Figure 1 here]

Our Washington site (Figure 1) is located in the northwest part of Washington state and includes the North Cascades National Park and part of the Mount Baker-Snoqualmie National Forest. Elevation ranges from 70 m in the southwest to 3300 m in the northeast. The North Cascade Range is oriented in a NW-SE direction and divides the region into distinct regimes; cool and wet to the west of the range during winter and cold and dry to the east. Summers are typically dry throughout the region. The predominant vegetation is evergreen forest, which covers more than 60% of the site area. Major tree species include Western Hemlock, Pacific Silver Fir, Subalpine Mountain Hemlock, Alpine, Subalpine Fir, and Douglas Fir [*Crawford et al.*, 2009]. Other significant vegetation are shrubs, covering 14% of the territory, grasslands are

8% of the area, and deciduous forests are 1% of the area. The most distinctive feature of this landscape is the arc of perennial snow that covers about 1.5% of the land area. Soils are predominantly Andisols, Inceptisols, and exposed rock formation (rock outcrops) at higher altitudes of the mountain range [*Wolock*, 1997]. Rock outcrops account for almost 8% of the area. Annual mean temperature ranges from -4.9°C at higher altitudes to 10.5°C at lower altitudes. Annual precipitation varies between 460 mm, east of the Cascade Mountains, to 2087 mm on the western side of the mountains [*Hijmans et al.*, 2005]. The total area covered is 9,400 km^2.

2.2 Data

Table 1 summarizes the spatial data used as model inputs at each of the two study sites. We use the 32 day raw composite satellite images from Landsat 7, specifically seeking information on the values of spectral bands 1, 2, 3, and 4, which correspond to the blue, green, red, and near infrared colour channels, respectively. Both historical and future climate datasets (items 4 and 5 in Table 1) are obtained from the WorldClim dataset [*Hijmans et al.*, 2005]. The historical observed climate data by *Hijmans et al.* [2005] has used observed meteorological station data (47544 stations for precipitation and 24542 stations for air temperature) from a variety of sources, such as Global Historical Climatology Network (GHCN v2), World Meteorological Organization's Climate Normals (WMO CLINO), and Food and Agricultural Organization's Agroclimatic Database (FAOCLIM 2.0). These observed point data have been interpolated over a 1 km global grid using the thin-plate smoothing spline algorithm. As shown in Figure 1 of *Hijmans et al.* [2005], the meteorological station density is amongst the highest in the continental United States. The downscaled 2061-2080 climate data from the GISS E2 model output are for the Representative Concentration Pathway (RCP) 8.5 scenario [*IPCC*, 2013]. To ensure fast computation as well as uniformity amongst the different datasets, we resample all the above-mentioned data onto a common 150 m resolution grid.

[Insert Table 1 here]

3 Methods

3.1 Machine learning model

We use the RF model [*Breiman*, 2001], which is an ensemble-based machine learning method, to predict the spectral band information of Landsat images. The spatial data used as model inputs, i.e., the predictor variables, are elevation, aspect, slope, mean annual precipitation and temperature. The model outputs, i.e., the target variables, are the spectral values from bands 1, 2, 3, and 4 of the Landsat image.

Each ensemble member in the RF model is a Decision Trees (DT) model, which is essentially an inverted binary tree structure where splitting rules govern the flow of decisions. The DT algorithm begins at the top node and proceeds down through internal nodes and branches. There are two main types of DT models: (1) CT, which are used when the data type of target variables is categorical, and (2) RT, which are used when the data type of target variables is numerical. Since our target variables are the spectral band values in every pixel of the Landsat data, we use RT as the base ensemble constituent of our RF model. Each node of RT is a binary split that is conditional based on the value of a predictor variable. The particular form of RT that we use here is Classification and Regression Tree (CART). CART builds a RT in a top-down manner, first creating a root node and progressively splitting the data into two sub-trees. The final output of RF model is the mean of the output from all individual RT models in the ensemble.

A drawback of the DT models is their tendency to overfit the training dataset by building very deep trees [*Bramer*, 2007]. This can lead to poor model performance when making predictions outside the training dataset. RF models reduce the risk of overfitting in two main ways. Firstly, given that the RF model structure is an ensemble of a large number of DT models, its output is not overly dependent on that of any single DT model. Secondly, when creating the training dataset for its ensemble member models, the RF model uses the bootstrap aggregating method (also referred to as bagging) [*Breiman*, 1996]. In this method, the original training dataset is sampled with replacement, thereby creating a sub-sampled dataset that has the same length as the original training dataset. The use of bagging method ensures that: (1) each individual DT model in the ensemble is trained with a slightly different dataset, and (2) part of the original training dataset that is left out due to bagging can be used as the test dataset to determine model performance (also known as the out-of-bag (OOB) score). Here, we use the coefficient of determination (R^2) to measure the RF model's OOB score.

One of the main controlling factors in RF model's performance is the number of its ensemble members (i.e., individual RT). Typically, having too few ensemble members leads to a poor OOB score, and increasing the number of ensemble members can improve the OOB score. However, the improvement in model performance becomes marginal once a certain threshold of ensemble members is crossed, and having too many ensemble members simply adds to the computational cost without any performance gain. During preliminary tests of the RF model with our datasets, we found that having more than 100 RT model ensembles provides virtually no improvement in the OOB score. Therefore, for all the results presented in this paper, our RF model consists of an ensemble of 100 RT models.

During the training phase of RF model, we use the historical climate data (see Table 1) and topographic variables (elevation, slope, and aspect) as model inputs. The spectral band information from Landsat 7 images is used for comparison with model outputs to calibrate the RF model. In the prediction phase, the RF model uses the future climate data and the topographic variables as model inputs. We use the RF model in the scikit-learn machine learning package that is implemented in Python® programming language [*Pedregosa et al.*, 2011].

3.2 Post-processing of the model outputs

The output of RF model is the spectral band information of the Blue, Green, Red, and Near Infrared bands of the Landsat image. We use this output information to create two derived products: (1) a true colour photorealistic image consisting of the Red, Green and Blue (RGB) colour bands, and (2) an image showing the NDVI values of the study sites. NDVI value for each pixel is calculated using the following formula.

$$\mathrm{NDVI} = \frac{B_4 - B_3}{B_4 + B_3} \qquad (1)$$

where, B_4 is the Near Infrared colour band and B_3 is the Red colour band of a Landsat 7 satellite image.

In addition to the OOB score obtained during the RF model's training phase (see Section 3.1), we calculate two more error metrics to assess the model performance for the final trained images. For the photorealistic image, the error at each pixel is calculated as follows.

$$E_{\mathrm{RGB}} = \sqrt{\left(B_{1,\mathrm{obs}} - B_{1,\mathrm{pred}}\right)^2 + \left(B_{2,\mathrm{obs}} - B_{2,\mathrm{pred}}\right)^2 + \left(B_{3,\mathrm{obs}} - B_{3,\mathrm{pred}}\right)^2} \qquad (2)$$

where, B_1 is the Blue colour band, B_2 is the Green colour band, and obs and pred denote the observed and model predicted spectral band values, respectively. For the NDVI image, the error at each pixel is calculated as follows.

$$E_{\mathrm{NDVI}} = \mathrm{NDVI}_{\mathrm{obs}} - \mathrm{NDVI}_{\mathrm{pred}} \qquad (3)$$

4 Results and Discussion

We first present the results from the RF model's training phase, which uses the topographic and historical climate data to train the model for predicting the four spectral band values (Blue, Green, Red, and Near Infrared) of the Landsat image. For the New Mexico site, the OOB R^2 value for the prediction of four spectral band values is 0.79. For the Washington site, the OOB R^2 value is 0.73. Figure 2 compares the original Landsat and the trained true colour photorealistic images for both study sites. Images produced using the RF model are able to capture almost all the major land cover features at both sites, and there is good visual agreement with the original Landsat images.

[Insert Figure 2 here]

Figure 3 compares the NDVI values between the original and trained images at both study sites. For the New Mexico site, the R^2 value between observed and simulated NDVI values is 0.97. For the Washington site, $R^2 = 0.96$ between the observed and simulated NDVI values. It is worth noting here that the R^2 values are much higher for NDVI because at each site we compare all the pixels between the observed and simulated data, whereas for the raw spectral band values, we only compare the pixels that were left out from training due to bagging.

[Insert Figure 3 here]

Figure 4 shows the RGB error between the original Landsat and the model generated photorealistic images calculated at each pixel using Equation 2. The error across RGB band values is lower at the New Mexico site, where there is no prominent geographical pattern for high error values. Conversely, the Washington site has higher error across the RGB band values, and the high error pixels are predominantly located in areas adjacent to the perennial snow cover. Figure 5 shows the error between the original and model generated NDVI images calculated at each pixel using Equation 3. Consistent with the RGB error shown in Figure 4, the NDVI error values are lower at the New Mexico site compared to the Washington site.

[Insert Figure 4 here]

[Insert Figure 5 here]

Next, we focus on the prediction phase of the RF model, which uses the topographic and future climate data (see Table 1) to predict the spectral band values for the RCP 8.5 climate change scenario. Figure 6 compares the historical (trained) and the future (predicted) true colour photorealistic images for both study sites. For the New Mexico site, the most prominent change is the increase in vegetation cover within the forested areas on either side of the Rio Grande river. For the Washington site, there is a substantial decrease in the perennial snow cover in the vicinity of Mount Baker (top left of the image) as well as across other mountainous areas along the Cascades Mountain Range. Many areas that appear as snow covered in the trained historical image are replaced by bare ground in the future scenario image. Figure 7 shows the NDVI images at both study sites for the historical (trained) and the future (predicted) scenarios. The overall increase in vegetation cover at the New Mexico site is discernible from the NDVI comparison. Interestingly, the reduction in perennial snow cover for the Washington site can be perceived through the increase in NDVI values in the mountainous areas.

[Insert Figure 6 here]

[Insert Figure 7 here]

We have attempted to demonstrate that a machine learning model that is trained to predict the spectral band information of satellite images can be highly useful for scenario-based assessment of future land cover. Moreover, given the richness of information available from spectral band values, it is possible to create several derived products to analyse (and visualize) land cover response to climate change from multiple perspectives. In our view, this is a non-trivial improvement from previous land cover change studies which had limited the application of machine learning models to categorical land cover classification data [*Rogan et al.*, 2008; *Schneider*, 2012; *Pearson et al.*, 2013]. It is worth mentioning here that the categorical land cover classification data itself is a product that is derived from satellite image data, similar to the photorealistic images and NDVI data shown in our study. Several methods, many of them based on machine learning, exist to convert the satellite's spectral band information into land cover classes [*Friedl and Brodley*, 1997; *DeFries and Chan*, 2000; *Hansen et al.*, 2000; *Qian et al.*, 2015]. We would also like to note that our focus on predicting only the first four spectral bands of the Landsat 7 images was governed by our choice of derivative products, the NDVI and RGB images (which require the use of first four bands only). Nonetheless, the techniques presented in this study are applicable to predicting the information from any desired number of satellite spectral bands, depending on the final product sought by the end user.

Our preference for choosing a RF machine learning model in this study was partly due to the fact that its ensemble constituents are comprised of DT models, which offers a number of attractive features over other statistical learning techniques. DT models are non-parametric and therefore make no assumptions regarding the distribution of the data. They are structurally explicit models and provide for a clear interpretation of the connections between the predictor and target variables. Normalization of attribute distances is unnecessary in these models, and their internal structure (essentially a cascading set of data splitting decisions) makes them much more tolerant to redundancies in the information content among the input variables [*Song and Lu*, 2015]. In addition, they tend to be computationally faster than other machine learning techniques [*Witten and Frank*, 2005; *Kotsiantis et al.*, 2007; *Rogan et al.*, 2008; *Schneider*, 2012] such as NN or BRT, and certainly faster than the physically based mechanistic models for a similar resolution data and areal extent. Lastly, as we had mentioned in Section 3.1, the ensemble averaging process in a RF model mitigates the drawbacks caused by the direct use of a standalone DT model. Nonetheless, there are a few assumptions and limitations built into our methodology. Firstly, our model requires long term climatic averages of precipitation and air temperature as inputs. These were chosen because the development of natural vegetation cover is a gradual process and would be a function of past climate over a long time period (in the order of decades) [*Dale*, 1997; *Kangur et al.*, 2005; *Soudzilovskaia et al.*, 2013], especially for forested areas which are abundant in both our study sites. Unfortunately, this makes the model unsuitable for change detection at short time scales, and a time gap of several decades would be needed between the training and prediction dataset to obtain meaningful change detection. Secondly, our input data was resampled to a common grid resolution of 150 m prior to running the model, which was done to limit the computational expenditure in the desktop runtime setting. Grid resampling does bring another source of uncertainty to the model, but is unavoidable due to different resolutions of our input datasets. Nonetheless, it would be possible to run our model at finer spatial resolutions if additional computational resources are available to the user.

As we look forward, the method presented in our study offer both challenges and opportunities. Firstly, our model presumes that the land cover change for the 2061-2080 period is simply the application of learned rules from the historical period to the climate changed environment. Many sites within our two study regions have experienced disturbance due to, for example, grazing pressure and fires [*Everett et al.*, 2000; *Floyd et al.*, 2003; *Allen*, 2007]. However, to a large extent, this is mitigated by the fact that our land cover training is conducted over regions that are much larger than the scale of a typical disturbance. Secondly, the predicted land cover for 2061-2080 period does not indicate the velocity of land cover change in response to changes in precipitation and air temperature [*Loarie et al.*, 2009]. Thus, our model does not provide any mechanistic understanding of how the final predicted state of land cover will be reached.

Within the limits of these challenges, the method presented here does provide a few opportunities. Monthly Landsat images are available at the 16 and 32 day time frames going back to 2002, and can provide ample raw data to explore how the seasonality of vegetation will be altered in a future scenario. Ongoing improvements in the satellite sensor technology, such as those in the recently launched Landsat 8 satellite [*Knight and Kvaran*, 2014; *Roy et al.*, 2014], also have the potential to provide increasingly better quality input data to land cover change models. The fast computational speed of the machine learning models permit the rendering of future land cover over much larger areas than our study regions, possibly even covering the entire continental USA. The five predictor variables we used were obtained from three primary data sources: rainfall, air temperature, and elevation (slope and aspect are derivative products of elevation), and were chosen based on what we judged to be important factors for predicting land cover. Nonetheless, we cannot rule out the possibility that, at least in some regions, inclusion of different types of predictor variables could improve the machine learning model's capability to predict land cover. Therefore, there is opportunity to experiment with the predictor variables by adding to or modifying the data sources.

5 Conclusions

In this paper, our goal was to extend the utility of machine learning based land cover change models to predict the spectral band information of satellite based land cover images. We used the topographic and historical climate data from two large sites in the United States to train a RF machine learning model to predict the spectral values from bands 1 (Blue), 2 (Green), 3 (Red), and 4 (Near Infrared) of Landsat 7 image. We then used the trained model to explore what the land cover might look like for a climate change scenario during the 2061-2080 period through the two derived products. Our results showed that the RF model can accurately reproduce the land cover properties for historical data and is able to provide realistic rendering of future land cover for a climate change scenario. The two derived land cover products (photorealistic RGB image and NDVI image) shown in our results demonstrate that the direct prediction of spectral band information is helpful for deriving ecologically relevant products. We consider this a major strength of our proposed approach because it enables the analysis of land cover change from multiple perspectives.

What land cover change will occur over the next 100 years is highly uncertain. However, presuming little is done to reduce the rate of CO_2 emissions, the global air temperatures for the 2081–2100 period are projected to be to 1.5 - 4.8 °C higher than for the 1986–2005 period [*IPCC*, 2013]. This will almost certainly impact regional and global land cover [*Krinner et al.*, 2005; *Beer et al.*, 2007; *Sitch et al.*, 2008; *Anav et al.*, 2010; *Hickler et al.*, 2012]. We hope that the method presented here makes a useful contribution towards understanding and predicting these changes.

Acknowledgements

This work was supported by the National Science Foundation under Award Number 1027870 (CDI-Type Small Resources Supercomputing: High Performance Computing in the Earth Sciences).

References

Abdelnour, A., M. Stieglitz, F. Pan, and R. McKane (2011), Catchment hydrological responses to forest harvest amount and spatial pattern, *Water Resour. Res.*, *47*, W09521–W09521, doi:10.1029/2010WR010165.

Abdelnour, A., R. B. McKane, M. Stieglitz, F. Pan, and Y. Cheng (2013), Effects of harvest on carbon and nitrogen dynamics in a Pacific Northwest forest catchment, *Water Resour. Res.*, *49*(3), 1292–1313, doi:10.1029/2012WR012994.

Allen, C. D. (2007), Interactions Across Spatial Scales among Forest Dieback, Fire, and Erosion in Northern New Mexico Landscapes, *Ecosystems*, *10*(5), 797–808, doi:10.1007/s10021-007-9057-4.

Alpaydin, E. (2014), *Introduction to machine learning*, MIT press.

Anav, A., F. D'Andrea, N. Viovy, and N. Vuichard (2010), A validation of heat and carbon fluxes from high-resolution land surface and regional models, *J. Geophys. Res. Biogeosciences*, *115*(G4), G04016, doi:10.1029/2009JG001178.

Barger, N. N., S. R. Archer, J. L. Campbell, C. Huang, J. A. Morton, and A. K. Knapp (2011), Woody plant proliferation in North American drylands: A synthesis of impacts on ecosystem carbon balance, *J. Geophys. Res. Biogeosciences*, *116*(G4), n/a-n/a, doi:10.1029/2010JG001506.

Beer, C., W. Lucht, D. Gerten, K. Thonicke, and C. Schmullius (2007), Effects of soil freezing and thawing on vegetation carbon density in Siberia: A modeling analysis with the Lund-Potsdam-Jena Dynamic Global Vegetation Model (LPJ-DGVM), *Global Biogeochem. Cycles*, *21*(1), n/a-n/a, doi:10.1029/2006GB002760.

Bishop, C. M. (2006), *Pattern recognition and machine learning*, springer New York.

Bramer, M. (2007), *Principles of data mining*, Springer.

Breiman, L. (1996), Bagging predictors, *Mach. Learn.*, *24*(2), 123–140, doi:10.1007/BF00058655.

Breiman, L. (2001), Random Forests, *Mach. Learn.*, *45*(1), 5–32, doi:10.1023/A:1010933404324.

Brovkin, V., M. Claussen, E. Driesschaert, T. Fichefet, D. Kicklighter, M. F. Loutre, H. D. Matthews, N. Ramankutty, M. Schaeffer, and A. Sokolov (2006), Biogeophysical effects of historical land cover changes simulated by six Earth system models of intermediate complexity, *Clim. Dyn.*, *26*(6), 587–600, doi:10.1007/s00382-005-0092-6.

Brown de Colstoun, E. C., M. H. Story, C. Thompson, K. Commisso, T. G. Smith, and J. R. Irons (2003), National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier, *Remote Sens. Environ.*, *85*(3), 316–327, doi:http://dx.doi.org/10.1016/S0034-4257(03)00010-5.

Campbell, J. L., S. V Ollinger, G. N. Flerchinger, H. Wicklein, K. Hayhoe, and A. S. Bailey (2010), Past and projected future changes in snowpack and soil frost at the Hubbard Brook

Experimental Forest, New Hampshire, USA, *Hydrol. Process.*, *24*(17), 2465–2480, doi:10.1002/hyp.7666.

Crawford, R. C., C. B. Chappell, C. C. Thompson, and F. J. Rocchio (2009), *Vegetation classification of Mount Rainier, North Cascades, and Olympic National Parks. Plant association descriptions and identification keys*, Natural Resource Technical Report NPS/NCCN/NRTR—2009/D-586. US Department of the Interior, National Park Service, Natural Resource Program Centre, Fort Collins, CO, US.

Dale, V. H. (1997), The Relationship Between Land- Use Change and Climate Change, *Ecol. Appl.*, *7*(3), 753–769.

DeFries, R. S., and J. C.-W. Chan (2000), Multiple Criteria for Evaluating Machine Learning Algorithms for Land Cover Classification from Satellite Data, *Remote Sens. Environ.*, *74*(3), 503–515, doi:http://dx.doi.org/10.1016/S0034-4257(00)00142-5.

Domingos, P. (2012), A Few Useful Things to Know About Machine Learning, *Commun. ACM*, *55*(10), 78–87, doi:10.1145/2347736.2347755.

Everett, R. L., R. Schellhaas, D. Keenum, D. Spurbeck, and P. Ohlson (2000), Fire history in the ponderosa pine/Douglas-fir forests on the east slope of the Washington Cascades, *For. Ecol. Manage.*, *129*(1–3), 207–225, doi:http://dx.doi.org/10.1016/S0378-1127(99)00168-1.

Floyd, M. L., T. L. Fleischner, D. Hanna, and P. Whitefield (2003), Effects of Historic Livestock Grazing on Vegetation at Chaco Culture National Historic Park, New Mexico, *Conserv. Biol.*, *17*(6), 1703–1711, doi:10.1111/j.1523-1739.2003.00227.x.

Foster, J. L. (1989), The Significance of the Date of Snow Disappearance on the Arctic Tundra as a Possible Indicator of Climate Change, *Arct. Alp. Res.*, *21*(1), 60–70, doi:10.2307/1551517.

Foster, J. L., J. W. Winchester, and E. G. Dutton (1992), The date of snow disappearance on the Arctic tundra as determined from satellite, meteorological station and radiometric in situ observations, *Geosci. Remote Sensing, IEEE Trans.*, *30*(4), 793–798, doi:10.1109/36.158874.

Friedl, M. A., and C. E. Brodley (1997), Decision tree classification of land cover from remotely sensed data, *Remote Sens. Environ.*, *61*(3), 399–409, doi:http://dx.doi.org/10.1016/S0034-4257(97)00049-7.

Groisman, P. Y., T. R. Karl, and R. W. Knight (1994), Observed Impact of Snow Cover on the Heat Balance and the Rise of Continental Spring Temperatures, *Sci.* , *263*(5144), 198–200, doi:10.1126/science.263.5144.198.

Guisan, A., and N. E. Zimmermann (2000), Predictive habitat distribution models in ecology, *Ecol. Modell.*, *135*(2–3), 147–186, doi:http://dx.doi.org/10.1016/S0304-3800(00)00354-9.

Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J. M. C. C. Overton, R. Aspinall, and T. Hastie (2006), Making better biogeographical predictions of species' distributions, *J. Appl. Ecol.*, *43*(3), 386–392, doi:10.1111/j.1365-2664.2006.01164.x.

Hall, D. K., C. J. Crawford, N. E. DiGirolamo, G. A. Riggs, and J. L. Foster (2015), Detection of earlier snowmelt in the Wind River Range, Wyoming, using Landsat imagery, 1972–2013, *Remote Sens. Environ.*, *162*, 45–54, doi:http://dx.doi.org/10.1016/j.rse.2015.01.032.

Hansen, M. C., R. S. Defries, J. R. G. Townshend, and R. Sohlberg (2000), Global land cover classification at 1 km spatial resolution using a classification tree approach, *Int. J. Remote Sens.*, *21*(6–7), 1331–1364, doi:10.1080/014311600210209.

Hickler, T. et al. (2012), Projecting the future distribution of European potential natural vegetation zones with a generalized, tree species-based dynamic vegetation model, *Glob. Ecol. Biogeogr.*, *21*(1), 50–63, doi:10.1111/j.1466-8238.2010.00613.x.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis (2005), Very high resolution interpolated climate surfaces for global land areas, *Int. J. Climatol.*, *25*(15), 1965–1978, doi:10.1002/joc.1276.

Homer, C. G., G. Xian, C. L. Aldridge, D. K. Meyer, T. R. Loveland, and M. S. O'Donnell (2015), Forecasting sagebrush ecosystem components and greater sage-grouse habitat for 2050: Learning from past climate patterns and Landsat imagery to predict the future, *Ecol. Indic.*, *55*, 131–145, doi:http://dx.doi.org/10.1016/j.ecolind.2015.03.002.

Im, J., and J. R. Jensen (2005), A change detection model based on neighborhood correlation image analysis and decision tree classification, *Remote Sens. Environ.*, *99*(3), 326–340, doi:http://dx.doi.org/10.1016/j.rse.2005.09.008.

IPCC (2013), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern - mental Panel on Climate Change*, Cambridge, United Kingdom and New York, NY, USA.

Kangur, A., H. Korjus, K. Jõgiste, and A. Kiviste (2005), A conceptual model of forest stand development based on permanent sample-plot data in Estonia, *Scand. J. For. Res.*, *20*(S6), 94–101.

Kelly, A. E., and M. L. Goulden (2008), Rapid shifts in plant distribution with recent climate change, *Proc. Natl. Acad. Sci.* , *105*(33), 11823–11826, doi:10.1073/pnas.0802891105.

Kerr, J. T., and M. Ostrovsky (2003), From space to species: ecological applications for remote sensing, *Trends Ecol. Evol.*, *18*(6), 299–305, doi:10.1016/S0169-5347(03)00071-5.

Klein, I., U. Gessner, and C. Kuenzer (2012), Regional land cover mapping and change detection in Central Asia using MODIS time-series, *Appl. Geogr.*, *35*(1–2), 219–234, doi:http://dx.doi.org/10.1016/j.apgeog.2012.06.016.

Knight, J. E., and G. Kvaran (2014), Landsat-8 Operational Land Imager Design, Characterization and Performance, *Remote Sens.* , *6*(11), doi:10.3390/rs61110286.

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas (2007), Supervised machine learning: A review of classification techniques,

Krinner, G., N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice (2005), A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem. Cycles*, *19*(1), n/a-n/a, doi:10.1029/2003GB002199.

Lenoir, J., J. C. Gégout, P. A. Marquet, P. de Ruffray, and H. Brisse (2008), A Significant Upward Shift in Plant Species Optimum Elevation During the 20th Century, *Sci.* , *320*(5884), 1768–1771, doi:10.1126/science.1156831.

Loarie, S. R., P. B. Duffy, H. Hamilton, G. P. Asner, C. B. Field, and D. D. Ackerly (2009), The velocity of climate change, *Nature*, *462*(7276), 1052–1055.

McIver, D. K., and M. A. Friedl (2002), Using prior probabilities in decision-tree classification of remotely sensed data, *Remote Sens. Environ.*, *81*(2–3), 253–261, doi:http://dx.doi.org/10.1016/S0034-4257(02)00003-2.

Mitchell, T. M. (1997), *Machine learning*, McGraw Hill, Burr Ridge, IL.

Mote, P. W. (2003), Trends in snow water equivalent in the Pacific Northwest and their climatic causes, *Geophys. Res. Lett.*, *30*(12), n/a-n/a, doi:10.1029/2003GL017258.

Parker, D. C., S. M. Manson, M. A. Janssen, M. J. Hoffmann, and P. Deadman (2003), Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review, *Ann. Assoc. Am. Geogr.*, *93*(2), 314–337, doi:10.1111/1467-8306.9302004.

Pauleit, S., R. Ennos, and Y. Golding (2005), Modeling the environmental impacts of urban land use and land cover change—a study in Merseyside, UK, *Landsc. Urban Plan.*, *71*(2–4), 295–310, doi:http://dx.doi.org/10.1016/j.landurbplan.2004.03.009.

Pearson, R. G., and T. P. Dawson (2003), Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful?, *Glob. Ecol. Biogeogr.*, *12*(5), 361–371, doi:10.1046/j.1466-822X.2003.00042.x.

Pearson, R. G., S. J. Phillips, M. M. Loranty, P. S. A. Beck, T. Damoulas, S. J. Knight, and S. J. Goetz (2013), Shifts in Arctic vegetation and associated feedbacks under climate change, *Nat. Clim. Chang.*, *3*(7), 673–677.

Pedregosa, F. et al. (2011), Scikit-learn: Machine Learning in {P}ython, *J. Mach. Learn. Res.*, *12*, 2825–2830.

Pettorelli, N., J. O. Vik, A. Mysterud, J.-M. Gaillard, C. J. Tucker, and N. C. Stenseth (2005), Using the satellite-derived NDVI to assess ecological responses to environmental change, *Trends Ecol. Evol.*, *20*(9), 503–510, doi:10.1016/j.tree.2005.05.011.

Pitman, A. J. et al. (2009), Uncertainties in climate responses to past land cover change: First results from the LUCID intercomparison study, *Geophys. Res. Lett.*, *36*(14), n/a-n/a, doi:10.1029/2009GL039076.

Qian, Y., W. Zhou, J. Yan, W. Li, and L. Han (2015), Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery, *Remote Sens.* , *7*(1), doi:10.3390/rs70100153.

Rogan, J., J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts (2008), Mapping land-cover modifications over large areas: A comparison of machine learning algorithms, *Remote Sens. Environ.*, *112*(5), 2272–2283, doi:http://dx.doi.org/10.1016/j.rse.2007.10.004.

Root, T. L., J. T. Price, K. R. Hall, S. H. Schneider, C. Rosenzweig, and J. A. Pounds (2003), Fingerprints of global warming on wild animals and plants, *Nature*, *421*(6918), 57–60.

Roughgarden, J., S. W. Running, and P. A. Matson (1991), What Does Remote Sensing Do For Ecology?, *Ecology*, *72*(6), 1918–1922, doi:10.2307/1941546.

Roy, D. P. et al. (2014), Landsat-8: Science and product vision for terrestrial global change research, *Remote Sens. Environ.*, *145*, 154–172, doi:http://dx.doi.org/10.1016/j.rse.2014.02.001.

Schneider, A. (2012), Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach, *Remote Sens. Environ.*, *124*, 689–704, doi:http://dx.doi.org/10.1016/j.rse.2012.06.006.

Sitch, S. et al. (2003), Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, *Glob. Chang. Biol.*, *9*(2), 161–185, doi:10.1046/j.1365-2486.2003.00569.x.

Sitch, S. et al. (2008), Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs), *Glob. Chang. Biol.*, *14*(9), 2015–2039, doi:10.1111/j.1365-2486.2008.01626.x.

Song, Y., and Y. Lu (2015), Decision tree methods: applications for classification and prediction, *Shanghai Arch. Psychiatry*, *27*(2), 130–135, doi:10.11919/j.issn.1002-0829.215044.

Soudzilovskaia, N. A., T. G. Elumeeva, V. G. Onipchenko, I. I. Shidakov, F. S. Salpagarova, A. B. Khubiev, D. K. Tekeev, and J. H. C. Cornelissen (2013), Functional traits predict relationship between plant abundance dynamic and long-term climate warming, *Proc. Natl. Acad. Sci.*, *110*(45), 18180–18184.

Stone, R. S., E. G. Dutton, J. M. Harris, and D. Longenecker (2002), Earlier spring snowmelt in northern Alaska as an indicator of climate change, *J. Geophys. Res. Atmos.*, *107*(D10), ACL 10-1-ACL 10-13, doi:10.1029/2000JD000286.

Stow, D. A. et al. (2004), Remote sensing of vegetation and land-cover change in Arctic Tundra Ecosystems, *Remote Sens. Environ.*, *89*(3), 281–308, doi:http://dx.doi.org/10.1016/j.rse.2003.10.018.

Tague, C., L. Seaby, and A. Hope (2009), Modeling the eco-hydrologic response of a Mediterranean type ecosystem to the combined impacts of projected climate change and altered fire frequencies, *Clim. Change*, *93*(1–2), 137–155, doi:10.1007/s10584-008-9497-7.

VanDerWal, J., H. T. Murphy, A. S. Kutt, G. C. Perkins, B. L. Bateman, J. J. Perry, and A. E. Reside (2013), Focus on poleward shifts in species' distribution underestimates the fingerprint of climate change, *Nat. Clim. Chang.*, *3*(3), 239–243.

Vapnik, V. N. (1999), An overview of statistical learning theory, *Neural Networks, IEEE Trans.*, *10*(5), 988–999, doi:10.1109/72.788640.

Verburg, P. H., K. Neumann, and L. Nol (2011), Challenges in using land use and land cover data for global change studies, *Glob. Chang. Biol.*, *17*(2), 974–989, doi:10.1111/j.1365-2486.2010.02307.x.

Walther, G.-R., E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg, and F. Bairlein (2002), Ecological responses to recent climate change, *Nature*, *416*(6879), 389–395.

Witten, I. H., and E. Frank (2005), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Witten, I. H., S. J. Cunningham, G. Holmes, R. McQueen, and L. Smith (1993), *Practical machine learning and its application to problems in agriculture*, University of Waikato, Department of Computer Science.

Wolock, D. M. (1997), *STATSGO soil characteristics for the conterminous United States*, US Geological Survey.

Tables

Table 1: Summary information of all the input data used for training the machine learning model

Attribute	**Source**	**Resolution**
Elevation	USGS National Elevation Dataset	30 m
Aspect	Calculated from Elevation data	30 m
Slope	Calculated from Elevation data	30 m
Historical mean annual temperature and precipitation	Worldclim – Normal 1950-2000 period [*Hijmans et al.*, 2005]	1000 m
Future mean annual temperature and precipitation	Worldclim – Downscaled GISS E2 2061-2080 period [*Hijmans et al.*, 2005]	1000 m
Landsat 7 reflectance imagery	For New Mexico: 16 October 1999 – 17 November 1999 For Washington: 12 July 2001 – 13 August 2001	30 m

Figures

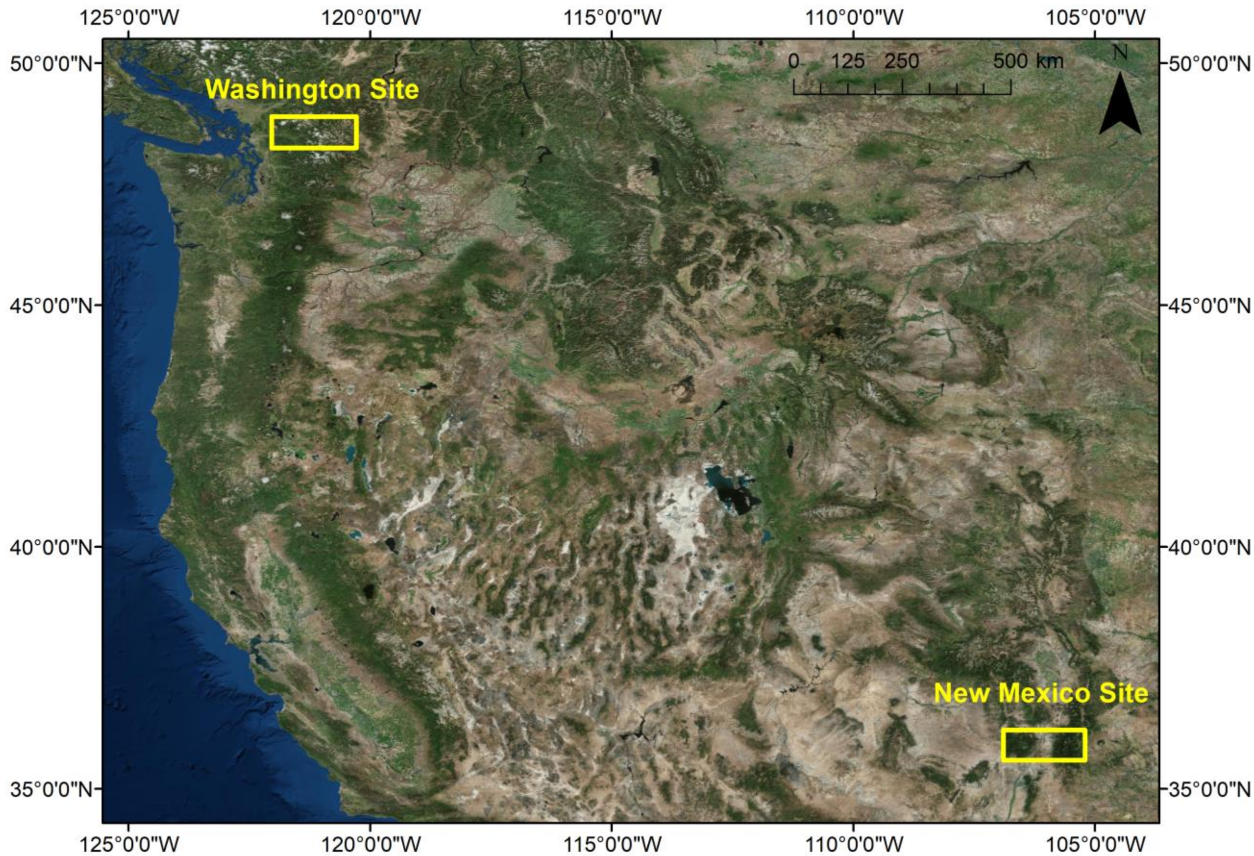

Figure 1: Location map of the two study sites.

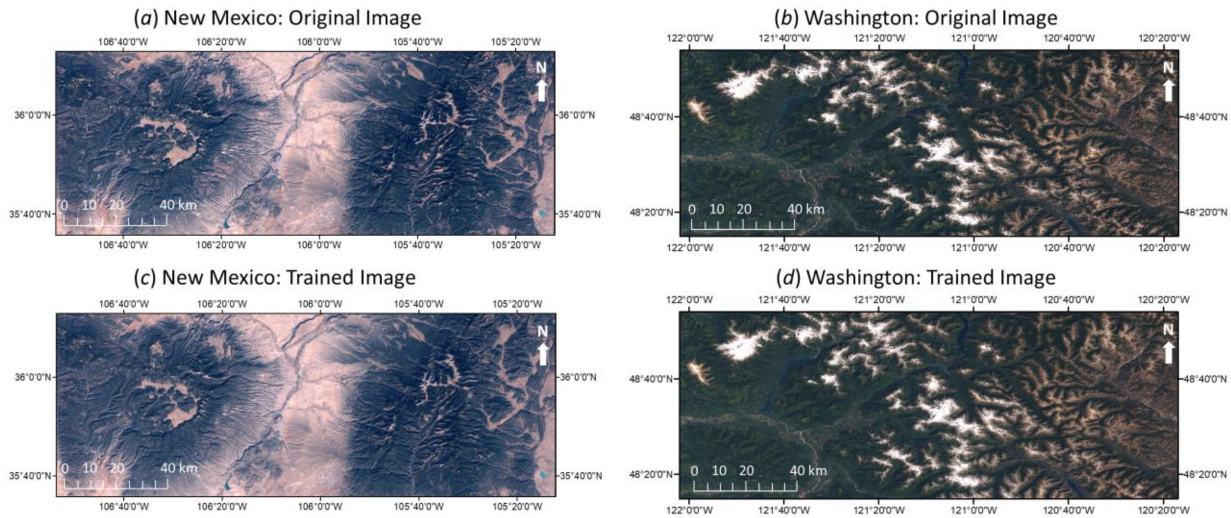

Figure 2: Comparison of the original Landsat 7 images and the RF model trained true colour photorealistic images for the two study sites.

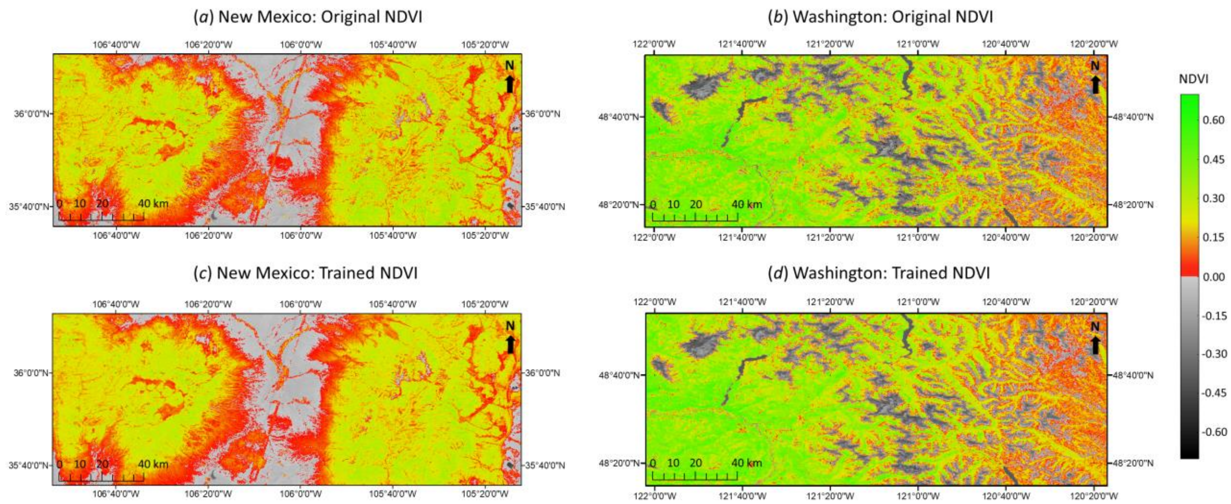

Figure 3: Comparison of the NDVI values between the original historical images (derived from Landsdat 7 using Equation 1) and the RF model trained images for the two study sites.

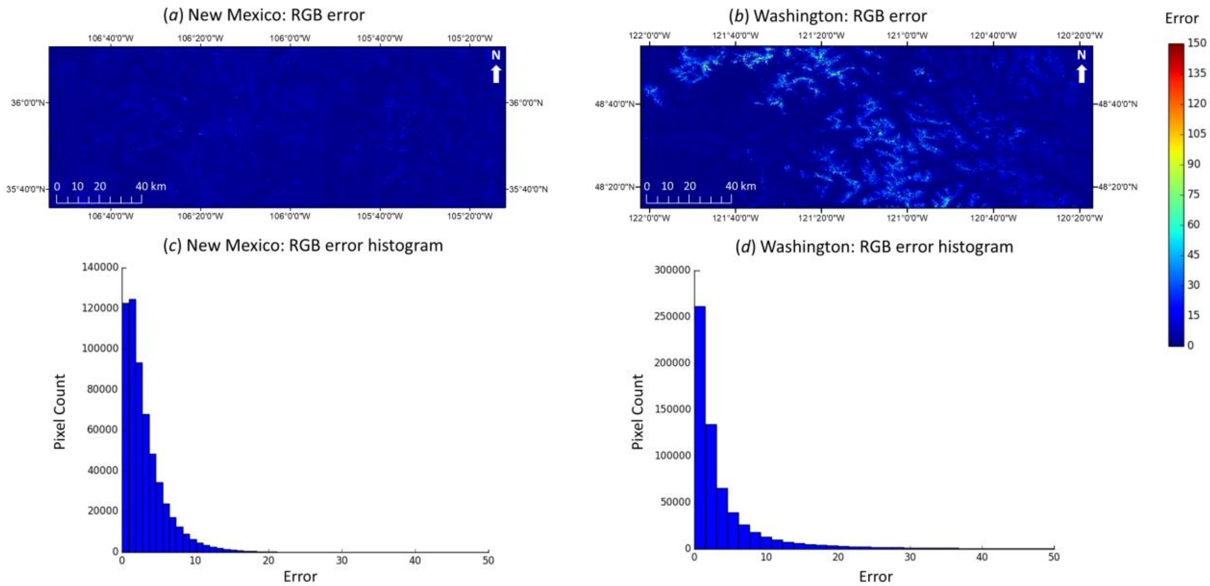

Figure 4: Spatial distribution of the error in RGB band values calculated for each pixel at (*a*) New Mexico site, and (*b*) Washington site. Also shown are the error histograms using the data from all the pixels at (*c*) New Mexico site, and (*d*) Washington site.

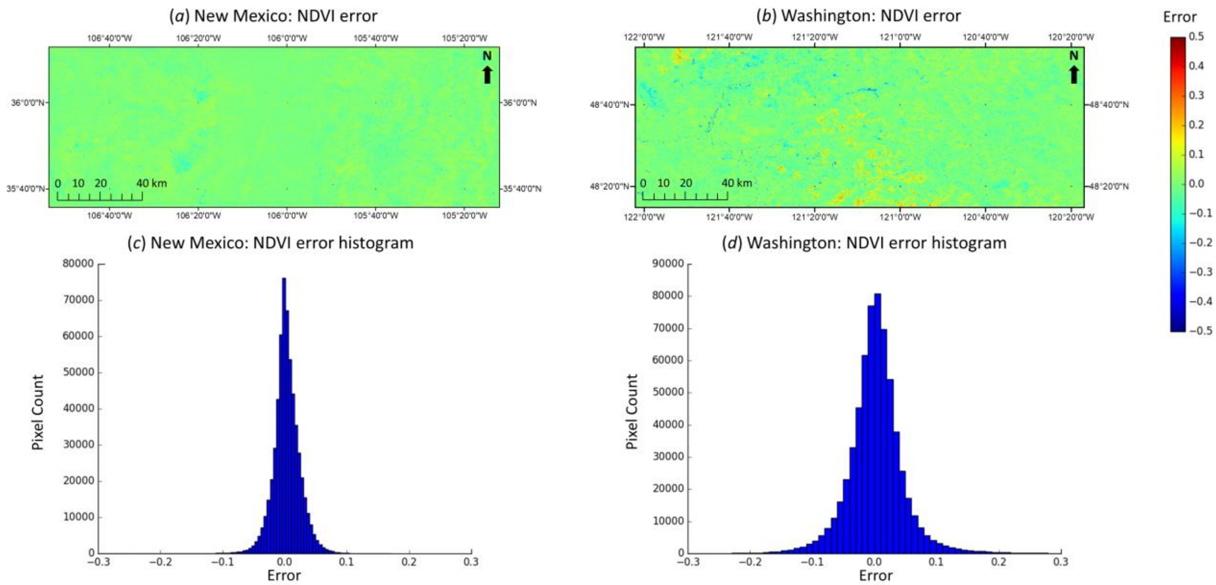

Figure 5: Spatial distribution of the error in NDVI values calculated for each pixel at (*a*) New Mexico site, and (*b*) Washington site. Also shown are the error histograms using the data from all the pixels at (*c*) New Mexico site, and (*d*) Washington site.

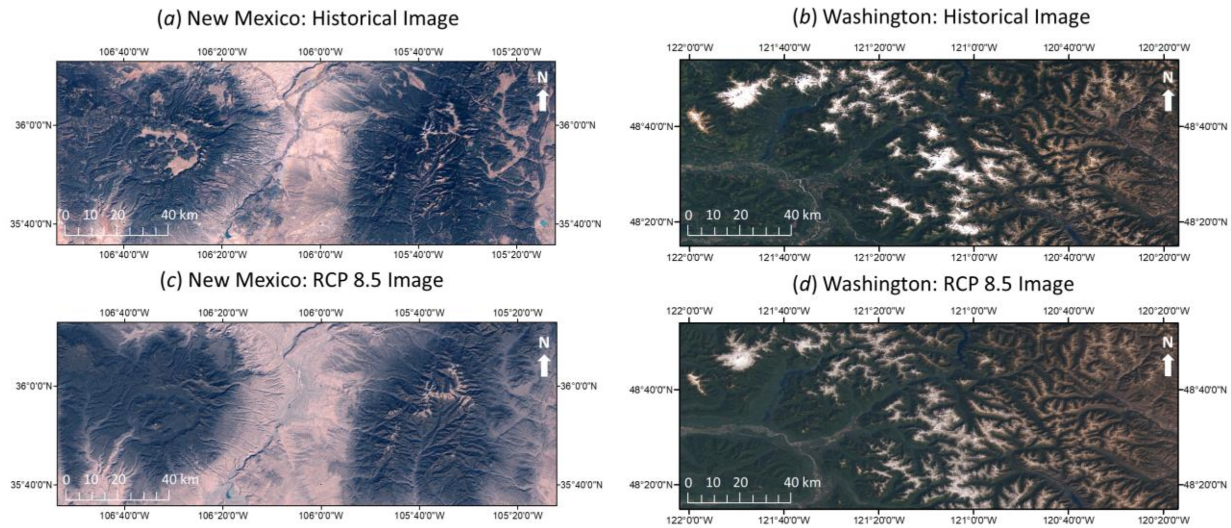

Figure 6: Comparison of the historical (RF model trained) and future (RF model predicted for RCP 8.5 scenario) true colour photorealistic images for the two study sites.

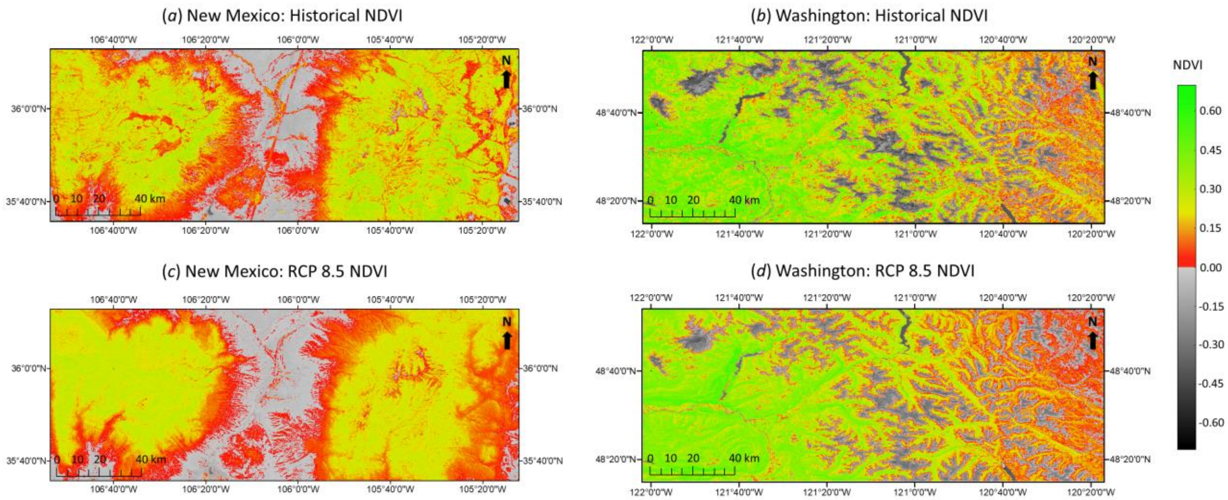

Figure 7: Comparison of the NDVI values between the historical (RF model trained) and future (RF model predicted for RCP 8.5 scenario) images for the two study sites.