

Multiple-scale prediction of forest loss risk across Borneo

Samuel A. Cushman · Ewan A. Macdonald · Erin L. Landguth ·
Yadvinder Malhi · David W. Macdonald

Received: 3 April 2016 / Accepted: 7 April 2017 / Published online: 18 May 2017
© The Author(s) 2017. This article is an open access publication

Abstract

Context The forests of Borneo have among the highest biodiversity and also the highest forest loss rates on the planet.

Objectives Our objectives were to: (1) compare multiple modelling approaches, (2) evaluate the utility of landscape composition and configuration as predictors, (3) assess the influence of the ratio of forest loss and persistence points in the training sample, (4) identify the multiple-scale drivers of recent forest loss and (5) predict future forest loss risk across Borneo.

Methods We compared random forest machine learning and logistic regression in a multi-scale approach to model forest loss risk between 2000 and

2010 as a function of topographical variables and landscape structure, and applied the highest performing model to predict the spatial pattern of forest loss risk between 2010 and 2020. We utilized a naïve model as a null comparison and used the total operating characteristic AUC to assess model performance.

Results Our analysis produced five main results. We found that: (1) random forest consistently outperformed logistic regression and the naïve model; (2) including landscape structure variables substantially improved predictions; (3) a ratio of occurrence to non-occurrence points in the training dataset that does not match the actual ratio in the landscape biases the predictions of both random forest and logistic regression; (4) forest loss risk differed between the three nations that comprise Borneo, with patterns in Kalimantan highly related to distance from the edge of the previous frontier of forest loss, while Malaysian Borneo showed a more diffuse pattern related to the structure of the landscape; (5) we predicted continuing

Electronic supplementary material The online version of this article (doi:[10.1007/s10980-017-0520-0](https://doi.org/10.1007/s10980-017-0520-0)) contains supplementary material, which is available to authorized users.

Samuel A. Cushman and Ewan A. Macdonald contributed equally to this work and are jointly first author.

S. A. Cushman
Rocky Mountain Research Station, United States Forest
Service, 2500 S. Pine Knoll Dr., Flagstaff, AZ 86001,
USA

E. A. Macdonald (✉) · Y. Malhi
Environmental Change Institute, School of Geography
and the Environment, University of Oxford, South Parks
Road, Oxford OX1 3QY, UK
e-mail: ewan.macdonald@zoo.ox.ac.uk

E. L. Landguth
Division of Biological Sciences, University of Montana,
Missoula, MT 59812, USA

E. A. Macdonald · D. W. Macdonald
Wildlife Conservation Research Unit, Zoology
Department, The Recanati-Kaplan Centre, University of
Oxford, Tubney House, Abingdon Road, Tubney,
Abingdon OX13 5QL, UK

very high rates of forest loss in the 2010–2020 period, and produced maps of the expected risk of forest loss across the full extent of Borneo.

Conclusions These results confirm that multiple-scale modelling using landscape metrics as predictors in a random forest modelling framework is a powerful approach to landscape change modelling. There is immense immanent risk to Borneo's forests, with clear spatial patterns of risk related to topography and landscape structure that differ between the three nations that comprise Borneo.

Keywords Borneo · Forest cover · Land-use change · Forest loss · Random forest · Logistic regression · Total operating characteristic

Introduction

Recent analysis of earth observation satellite data suggested that approximately 230 million ha of forest were lost due to disturbance globally in the period between 2000 and 2012 (c.19 million ha per year) with the greatest loss occurring in the tropics (Hansen et al. 2013). Rates of deforestation in South East Asia are high and accelerating, with Indonesia showing the largest recent increase in the rate of forest loss globally, while Malaysia has the highest level of forest loss in relation to land area (Hansen et al. 2013). Harris et al. (2012) estimate that 32% of carbon emissions from land-use change in the tropics arose in South and South-east Asia, and Indonesia was the second highest emitter, accounting for 13% of total CO₂ emissions from deforestation.

In light of these impacts and international commitments to reduce carbon emissions from deforestation and degradation (REDD+) in developing countries it is important to understand the dynamics of land-use change and predict which areas are at highest risk of forest loss. Multiple studies have attempted to quantify and predict future deforestation, most notably in the Amazon basin (Soares-Filho et al. 2006; Rosa et al. 2013), and in other forests around the world (Rideout et al. 2013; Vieilledent et al. 2013); however, to date we know of no analyses that predict future forest loss across the full extent of Borneo.

Published analyses of forest loss have employed a range of statistical techniques including logistic

regression (Chowdhury 2006; Echeverria et al. 2008), genetic algorithms (Venema et al. 2005; Soares-Filho et al. 2013), weights of evidence (Soares-Filho et al. 2010; Maeda et al. 2011) and cellular automata (Thapa et al. 2013). These models are based on a number of environmental and anthropogenic landscape features that are thought to drive deforestation rates. Commonly, these include distance to important features, as well as topography. Few studies have rigorously compared the performance of different approaches for modelling forest loss. To rigorously compare different modelling approaches in their performance in predicting forest loss Pontius et al. (2008) argue that the methods being compared must be applied to the same study landscape, but very few published studies have done this.

Logistic regression is almost certainly the modelling approach that has been most commonly used to predict forest loss (e.g., Chowdhury 2006; Echeverria et al. 2008), and has been the dominant method in multi-scale landscape modelling (McGarigal et al. 2016). Random forest (RF; Breiman 2001) is increasingly used in a range of applications including digital soil mapping (Grimm et al. 2008), forest biomass mapping (Baccini et al. 2012), species distribution modeling (Evans and Cushman 2009) and others given its often superior performance compared to other methods (Evans et al. 2011). RF is also gaining prominence in land-use classification (e.g., Aide et al. 2013; Grinand et al. 2013), where it outperforms classification and regression trees (CART; Rodriguez-Galiano et al. 2012) and maximum likelihood classifiers (Schneider 2012). Nonparametric procedures like RF are particularly effective at identifying complex multivariate associations, such as those that affect patterns of forest loss. Few analyses of forest loss have employed random forest (but see Aide et al. 2013; Grinand et al. 2013), and none to our knowledge quantitatively compare the performance of random forest to other commonly used modelling approaches, such as logistic regression, or utilize multi-scale optimization (*sensu* McGarigal et al. 2016).

Past applications of random forest have used several approaches for selecting the number of occurrence and non-occurrence cells used to train the model. For example, Chawla et al. (2003) and Chen et al. (2004) found that imbalance between the proportion of presence and absence classes can cause bias in the prediction and model-fit. They found that

when an imbalanced sample is present, the bootstrap of the data is biased towards the majority class, thus over-predicting the majority-class and under-predicting the minority. The resulting model fit can be deceptive, exhibiting very small overall error due to very small errors in the majority as a result of extremely high cross-classification error from the minority class. An alternative solution that is often used when there are many more absences than presences in a classification dataset is to shift the cutoff for the probability of present from the true ratio to 0.5 or something smaller (Evans and Cushman 2009). Given that unbiased predictions are essential for predictive models of forest loss to be reliable, it is critical to evaluate the effect of sample ratio and find the ratio that produces unbiased estimates of forest loss risk.

The idea that ecological patterns and processes interact across scales in space and time is a fundamental tenant in landscape ecology (Wiens 1989; Levin 1992). Despite this widespread recognition, a review of the current literature found that most habitat ecology papers fail to address multiple spatial pattern-process relationships, and less than 5% have optimized multi-scale relationships (McGarigal et al. 2016). While many recent studies of land-use change allude to spatial scales, they typically do so implicitly by including distance to various landscape features (roads, rivers, population center, etc.) as a metric (e.g., (Rideout et al. 2013; Rosa et al. 2013; Vieilledent et al. 2013), which is not a true multi-scale analysis (sensu McGarigal et al. 2016). To our knowledge there have been no multi-scale optimization efforts applied to landscape change modeling. Additionally, while several analyses consider some element of landscape composition (e.g., Rosa et al. 2013; Vieilledent et al. 2013; Rideout et al. 2013), we are aware of no studies that have considered both the composition and configuration of the multi-scale neighbourhood around a point as a predictor variable.

There are six main goals of this study. First, we evaluated the effects of varying the ratio of loss to persistence cells in the training data set on the predicted probability of forest loss in both logistic regression and Random Forest. Second, as suggested by Pontius et al. (2008), we sought to compare the performance of three modelling methods (random forest, logistic regression and a naïve model) by

applying them in the same landscape in the same time interval. Third, we sought to formally evaluate the utility of using landscape structure as predictors by computing models with and without landscape metrics. Fifth, we sought to identify variables and their operative scales that best predicted forest loss or forest persistence between 2000 and 2010 and compare them between the three nations that comprise Borneo (Kalimantan Indonesia, Malaysian Borneo and Brunei). Sixth, we sought to apply the best multi-scale models for each of these nations to predict the future risk of forest loss across Borneo in the 2010–2020 time period. We had several hypotheses. First, consistent with Chawala et al. (2003) and Chen et al. (2004) we expected that the probability map produced by random forest would under predict loss when the ratio of loss in the training sample was equal to that in the real landscape, in which persisting forest cells far outnumber forest loss cells. Second, we expected that random forest would outperform logistic regression and the naïve model in each nation. Third, we expected that models including landscape variables would have substantially higher performance than models that excluded them. Fourth, we predicted that risk of forest loss would be related to predictor variables similarly across the three nations. Fifth, we expected that forest loss risk would be related to (a) topography, with high risk at low elevations and flat areas, and low risk at higher elevations and steep terrain, (b) past forest loss in the surrounding region, (c) protected area status, and (d) human population parameters such as distance to large settlements and local population density, in that order of influence.

Methods

Study area

The analysis extent is the entire island of Borneo, with separate models constructed for Brunei, the Indonesian provinces of Kalimantan, and for Malaysian Borneo (combining the provinces of Sabah and Sarawak). The extent is 731,058 km².

Land cover change data for 2000–2010

We used land cover maps produced for the years 2000 and 2010 by the Centre for Remote Imaging, Sensing

and Processing (CRISP; Miettinen et al. 2011, 2012). These maps were based on MODIS surface reflectance product images (NASA 2010), Shuttle Radar Topography Mission (SRTM) 90 m version 4 digital elevation information (Jarvis et al. 2006) and several peatland distribution maps (Wahyunto and Subagio 2003, 2004; Selvaradjou et al. 2005; Wahyunto et al. 2006). Miettinen et al. (2011) used a three step classification process to classify the landscape into 13 categories: water, mangrove, peat swamp forest, lowland forest, lower montane forest, upper montane forest, plantation/regrowth, lowland mosaic, montane mosaic, lowland open, montane open, urban and large scale palm plantation. Accuracy assessment for these maps found overall accuracies of 83% for 2000 and 85% for 2010, with class accuracies between 75 and 85% for the non-forest classes and up to 97% for the forest classes (Miettinen et al. 2011, 2012). Overall, the accuracy assessment of these maps compares favourably to other studies (e.g., Clark et al. 2010: 79.3%; Friedl et al. 2010: 75%; Rozenstein and Karnieli 2011: 81%; Rodriguez-Galiano et al. 2012: 92%). Our analysis is based on combining all forest classes and non-forest classes into a binary forest/non-forest map. Peat swamp forest, lowland forest, lower montane forest and upper montane forest were reclassified to forest, while all other areas were considered to be non-forest for our analysis. Category aggregation (Aldwaik et al. 2015) for the 2010 land cover map presented in Miettinen et al. (2012) resulted in a classification where forest has a producer's accuracy of 93.2% and a user's accuracy of 91.9%, while non forest has a producer's accuracy of 93.9% and a user's accuracy of 94.9.

Response variable for analysis

The response variable in our analysis was a binary GIS layer depicting areas of the study area that were forest in 2000 and became non-forest or regenerating forest in 2010 (value of 1 or “forest loss”), and areas that were forest in 2000 and remained forest in 2010 (value of 0 or “persistence”). We produced maps of all pixels that were loss or persistence in the 2000–2010 time-period and used these as the source of the loss and persistence cells for model training and model assessment. Non Forest at 2000 was not part of the model training or assessment.

Predictor variables for analysis

A priori, we proposed several environmental and anthropogenic variables as predictors of recent forest loss (Table 1). Anthropogenic variables included: distance to large population centres, defined as areas with greater than 100 people/km²; regional population density, defined as the focal mean of population density within a 100 km radius; local population density defined as the point population density at the pixel scale of 250 m; and a map of designated protected areas.

Topographical variables included elevation from the ASTER DEM and several terrain complexity measures produced using the Geomorphometry and Gradient Metrics Toolbox (ArcGIS 10.0; Evans and Oakleaf 2012). These included: topographical roughness, which measures the topographical complexity of the landscape within a defined focal extent (Blaszczynski 1997; Riley et al. 1999), and relative slope position, which measures the relative position of the focal pixel within a defined extent on a gradient from valley bottom to ridge top (Evans et al. 2014). Given that topographical factors may be related to deforestation at a range of spatial scales (Wiens 1989), we calculated topographical roughness and relative slope position at six spatial extents including focal radii of 1, 10, 20, 30, 40 and 50 km. These were chosen because 1 km represents a single pixel and its very near neighbours, while the other scales span a range from local to regional neighbourhoods.

We also included FRAGSTATS metrics quantifying the extent and configuration of different land cover classes across a range of focal extents as predictor variables. The classes used in the analysis include: (1) water, (2) mangrove, (3) peat swamp forest, (4) lowland forest, (5) lower montane forest, (6) upper montane forest, (7) plantation or regrowth, (8) lowland mosaic, (9) montane mosaic, (10) lowland open, (11) montane open, (12) Urban (Miettinen et al. 2011, 2012). For each of these classes we used FRAGSTATS 4.0 (McGarigal et al. 2012) to calculate the percentage of the focal landscape covered by each class within the six focal window extents (1, 10, 20, 30, 40 and 50 km). In addition to class-level percentage of the landscape in each cover class, we calculated four landscape-level metrics quantifying the structure of the full multi-class mosaic to assess how landscape complexity was related to deforestation risk. These

Table 1 Predictor variables used in the analysis

Variable name	Description	Rational	Source
Anthropogenic variables			
Local population density	Local population density calculated as the point population density at the pixel scale of 250 m	Risk of forest loss hypothesised to be greater with increased local population due to increased demand and infrastructure (Carr 2004; Vieilledent et al. 2013)	CIESIN, CIAT (2005): Gridded Population of the World (year 2000), Version 3
Focal mean population density	Regional population density calculated as the focal mean of local population density within a 100 km radius	Risk of forest loss hypothesised to be greater with increased regional population due to increased demand and infrastructure (Carr 2004; Vieilledent et al. 2013)	CIESIN, CIAT (2005): Gridded Population of the World (year 2000), Version 3
Distance to large population centre	Distance to large population centre, defined as areas with more than 100 people per square kilometre	Risk of forest loss hypothesised to be greater with proximity to large population centres population due to increased demand and infrastructure (Carr 2004; Mon et al. 2012; Vieilledent et al. 2013)	CIESIN, CIAT (2005): Gridded Population of the World (year 2000), Version 3
Protected areas	Protected areas. A binary map with protected areas classified as 1 and other areas classified as 0	Risk of forest loss hypothesised to be lower within protected areas (although some studies have reported no statistical difference in deforestation between timber concessions and protected areas Gaveau et al. 2013)	IUCN, UNEP-WCMC (2014), The World Database on Protected Areas (year 2012)
Topographical variables			
Elevation	Point elevation from Aster DEM	Risk of forest loss hypothesised to decrease with elevation (Mon et al. 2012)	NASA LPDAAC (2011) Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model Version 2 (GDEM V2)
Roughness[x]	Focal topographical roughness within radius × km Values for × are: 1,10, 20, 30, 40, 50	Risk of forest loss hypothesised to decrease with increased topographic roughness at a moderate scale of 10–20 km	Geomorphology and Gradient Metrics Toolbox, http://evansmurphy.wix.com/evansspatial#!arcgis-gradient-metrics-toolbox/crro
Slope[x]	Focal relative slope position within radius × km Values for × are: 1,10, 20, 30, 40, 50	Risk of forest loss hypothesised to decrease with increased slope at a moderate scale of 10–20 km (Gaveau et al. 2013)	Geomorphology and Gradient Metrics Toolbox, http://evansmurphy.wix.com/evansspatial#!arcgis-gradient-metrics-toolbox/crro
Landscape metrics			
Landuse[x] (e.g. water[10])	Percentage of the landscape class within radius × km Landuse classes are: water, mangrove, peat swamp forest, lowland forest, lower montane forest, upper montane forest, plantation/regrowth, lowland mosaic, montane mosaic, lowland open, montane open, urban Values for × are: 1,10, 20, 30, 40, 50	Risk of forest loss hypothesised to increase with extent of previously deforested (not forest) cells area in the region surrounding the pixel. We expect that extremely local effects are probably minimal so the impact of previously deforested cells expected to be greatest at the scale of 10–30 km	FRAGSTATS analysis of reclassified Miettinen et al. (2011) land cover map

Table 1 continued

Variable name	Description	Rational	Source
Aggregation[x]	Aggregation Index for the full landscape mosaic within radius \times km Values for \times are: 1,10, 20, 30, 40, 50	Risk of forest loss hypothesised to decrease with increased aggregation. Low aggregation index represents a highly fragmented landscape that is potentially more susceptible to forest loss	FRAGSTATS analysis of reclassified Miettinen et al. (2011) land cover map
Edge[x]	Edge density for the full landscape mosaic within radius \times km Values for \times are: 1,10, 20, 30, 40, 50	Risk of forest loss hypothesised to increase with increased edge density. High edge density represents a highly fragmented landscape that is potentially more susceptible to forest loss	FRAGSTATS analysis of reclassified Miettinen et al. (2011) land cover map
Patch[x]	Patch Density for the full landscape mosaic within radius \times km Values for \times are: 1,10, 20, 30, 40, 50	Risk of forest loss hypothesised to increase with increased patch density. High patch density represents a highly fragmented landscape that is potentially more susceptible to forest loss	FRAGSTATS analysis of reclassified Miettinen et al. (2011) land cover map
Shannon[x]	Shannon's Diversity Index for the full landscape mosaic within radius \times km Values for \times are: 1,10, 20, 30, 40, 50	Risk of forest loss hypothesised to increase with increased Shannon's Diversity. High Shannon's diversity represents a highly fragmented landscape that is potentially more susceptible to forest loss	FRAGSTATS analysis of reclassified Miettinen et al. (2011) land cover map

landscape metrics were edge density (Edge[x]), patch density (Patch[x]), aggregation index (Aggregation[x]; which measures the compaction or aggregation of cover types), and Shannon's diversity index (Shannon[x]; which measures the diversity of the landscape mosaic within a window in terms of the richness and evenness of the cover types present). Each of these four metrics was also calculated for focal windows at the six spatial extents. FRAGSTATS metrics were calculated for the land-use maps in both the year 2000 and 2010 (Miettinen et al. 2011, 2012), the data from the year 2000 was used to calibrate the models while the 2010 data was used to predict the risk of future forest loss. All spatial layers were resampled to a 500 m pixel size and projected to an Albers conformal conic projection for analysis.

Modelling approaches

We used random forest machine learning and logistic regression to predict forest loss risk in the 2000–2010 time period from landscape conditions in 2000, separately for each nation within the study area

(Brunei, Malaysia, Indonesia), to assess how the drivers of deforestation may differ between nations. Random forest is a classification and regression tree (CART; De'ath and Fabricius 2000) based bootstrap method that corrects many of the known issues in CART, such as over-fitting (Breiman 2001; Cutler et al. 2007), and provides very well-supported predictions with large numbers of independent variables (Cutler et al. 2007). We used a modelling approach developed by Evans and Cushman (2009) to predict occurrence probabilities of deforestation for each nation using the random forest method (Breiman 2001; Cutler et al. 2007) as implemented in the package 'random Forest' (Liaw and Wiener 2002) in R (R Development Core Team 2008). We conducted a parallel set of analyses with logistic regression, for each nation using the same training data and predictor variable data set applied to the same landscape in the same time period as in the random forest models, enabling robust comparison of the performance of random forest and logistic regression.

We conducted the random forest and logistic regression analyses in two steps. First, we ran

univariate models across the multiple scales to identify the scale at which each variable had the strongest ability to predict forest loss in 2010 from the landscape condition in 2000, as suggested by McGarigal et al. (2016) as a robust approach for multi-scale model optimization. To accomplish this, we ran a series of single random forest analyses for each variable across the six scales in each nation and used the model improvement ratio (MIR; Murphy et al. 2010) to measure the relative predictive strength of each scale of the variable. The MIR calculates the permuted variable importance, represented by the mean decrease in out-of-bag error, standardized from zero to one. We compared the MIR scores for all scales for each variable, and retained the scale that had the highest MIR score for further multivariate modelling. In the logistic regression we used the same scales as identified in the random forest model to ensure the models were using identical input data, which is essential for a strict comparison of performance.

In the second step we used random forest and logistic regression to develop multivariate models predicting probability of forest loss in each nation in the 2000–2010 time period as a function of landscape condition across the suite of scale-optimized variables in the year 2000. To identify the most parsimonious random forest model for each nation we applied the Model Improvement Ratio (MIR; Murphy et al. 2010). In model selection using MIR, the variables are subset using 0.10 increments of MIR value, with all variables above the threshold retained for each model. This subset is always performed on the original model's variable importance to avoid over-fitting (Svetnik et al. 2004). We compared each subset model and selected the model that exhibits the lowest total out-of-bag error and lowest maximum within-class error. To identify the most parsimonious logistic regression model for each nation we used the same training and predictor variable set as in the random forest model for that nation, and employed all-subsets logistic regression with model averaging, using the “Dredge” function in the R package MuMIN (e.g., Timm et al. 2016; Chambers et al. 2016).

Model predictions for the random forest model were created by using a ratio of the majority votes-matrix to create a probability distribution. Random Forest makes predictions based on the plurality of votes across all bootstrap trees and not on a single rule set. This votes matrix can be scaled and treated as a probability given the error distribution of the model.

We used the function that (Evans and Cushman 2009) added to GridAsciiPredict (Crookston and Finley 2008) which uses the votes-probability function to write the probabilities to ASCII grid(s). Model predictions for the logistic regression model were created by calculating $p = \exp(z)/(1 + \exp(z))$, where p is the probability of forest loss, z is the linear combination of the model averaged coefficients multiplied by the independent variables.

Effects of training ratio of loss and persistence

We assessed the effects of varying the ratio of loss and persistence points by randomly choosing a relatively large number of persistence points for each nation and systematically varying the number of loss points, refitting the random forest and logistic regression models at each increment. For Kalimantan and Malaysian Borneo we chose 20,000 persistence cells, while in Brunei, which is much smaller, we chose 2311 persistence cells (25% of all persistence cells), and varied the proportion of loss cells at four intervals above and below the true ratio, rerunning the random forest and logistic regression predictor models at each ratio. To assess the effects of the ratio of loss to persistence cells in the training sample we calibrated the number of loss cells until the sum of the predicted probability map matched the observed number of loss cells in the nation over the 2000–2010 period. A simple simulation demonstrated that when the sum of predicted probability equals the number of actual loss cells, the prediction is unbiased in terms of the amount of predicted loss. We fit logarithmic regressions to the relationship between ratio of persistence and loss cells in the training dataset and bias in the predicted probability map for each nation.

Naïve model and model performance

Pontius et al. (2007) urges authors to compare any statistical landscape change model to a naïve model applied to the same landscape. Comparison to random allocation is not an appropriate naïve model because the factors driving forest loss are not random. For example, a naïve model could predict forest loss simply near previous forest loss. We constructed such a naïve model by making risk of forest loss inversely proportional to the distance from non forest at 2000 to the 10th power. This creates a nonlinear relationship

such that the risk of forest loss in the naïve model is 1 at the edge of non forest at 2000, but drops nonlinearly as one moves away from the edge into forest interior, reflecting the observed pattern that deforestation risk decreases faster than linearly with distance into existing forest (Fig. 1).

There are a multitude of ways to assess the performance of predictions of forest loss, and most previous studies have used the Kappa statistic (Cohen 1960) and similar measures of improvement of predicted classification compared to random assignment. However, following Pontius and Milones (2011), we eschewed the use of the Kappa statistic given that it does not report a meaningful statistical measure of predictive success, even when corrected to address the two different aspects of prediction related to predicted amount and predicted allocation (Pontius and Si 2014). In addition, the predictions produced by random forest and logistic regression are in the form of predicted probabilities, thus it is more meaningful to assess the continuous pattern or predicted probability in comparison to the actual observed changes than to cross-tabulate observed versus a single Boolean predicted change (Pontius and Si 2014). Cross tabulation requires use of a single threshold to transform the predicted probabilities into a Boolean response, which loses information concerning the various probabilities. We assessed the performance of the random forest and logistic regression predictions using area under the total operating characteristic curve, as suggested by Pontius and Si (2014) and Pontius and Parmentier (2014). We produced visualizations of the predicted probability of forest loss compared to the actual distribution of forest loss and persistence cells in which we overlaid the forest loss points on the predicted probability surface to visually display the association between predicted probabilities and observed changes.

Prediction to the 2010–2020 time period

Once random forest and logistic regression models had been produced for the 2000–2010 time period for each nation (Brunei, Malaysia, Indonesia), we applied these models to predict the risk of forest loss in the 2010–2020 time period by calculating the value of the predictor variables in the year 2010 and applying the models produced in the 2000–2010 time period to them. Having computed the predicted probability of forest loss in each cell over the 2010–2020 time

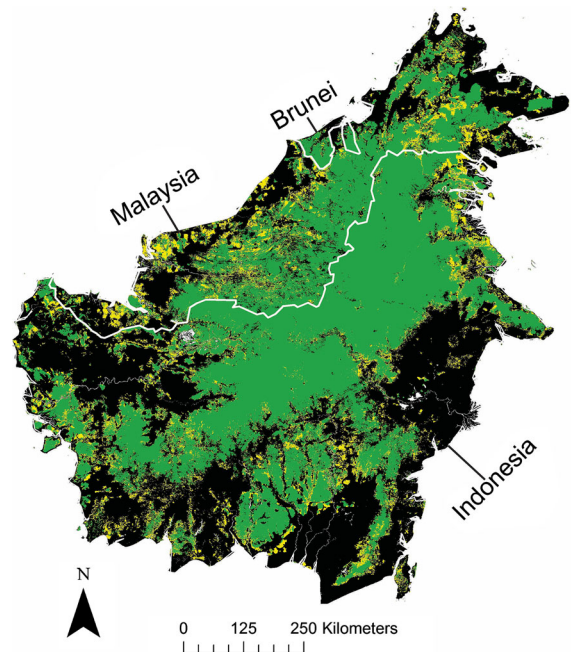


Fig. 1 Map of Borneo showing areas of forest loss between 2000 and 2010 in *yellow*, areas of forest persistence from 2000 to 2010 in *green*, and areas that were not forest in 2000 in *black*. (Color figure online)

period, we then calculated the proportion of each national territory in Borneo with forest loss risk greater than 25, 50 and 75% between 2010 and 2020.

Results

Model calibration

The calibrations of the number of loss to persist points used in the training sample for each of the three nations revealed a very clear result. In both the random forest calibration and the logistic regression calibration, the predicted amount of forest loss was biased upward when there were a higher proportion of loss points in the training sample than the actual landscape (Fig. 2). Conversely, predicted forest loss was biased downward when there were a lower proportion of deforested samples in the training data than in the actual landscape. An unbiased prediction of the amount of forest loss was obtained when the ratio of deforested to not-deforested points in the training sample matched that in the actual landscape, except in logistic regression when the actual proportion of deforested cells was

very low (Brunei), in which case logistic regression over-predicted forest loss at all ratios of deforested to not-deforested points. In addition, there was a very tight logarithmic relationship predicting the amount of bias (Fig. 2). In all three nations, the logistic regression calibration showed higher sensitivity to the ratio of loss and persist cells, with larger coefficient to the logarithmic equation leading to larger bias as the ratio departed the true ratio.

Observed and predicted forest loss amount in the 2000–2010 period

In the period between 2000 and 2010 the Miettinen et al. (2011) maps report that 5.1% of Brunei, 23.76%

of Malaysian Borneo and 15.44% of the forest present Kalimantan in 2000 was lost by 2010. The calibrated random forest model predicted risk surfaces for these nations that produced expected values of loss that were within $\frac{1}{2}$ of 1% of the actual observed amount of forest loss in each case (Table 2).

Model performance

Across the three nations, the calibrated random forest model out-performed the logistic regression and naïve models based on the TOC AUC value, which ranged from 0.931 in Malaysian Borneo, to 0.928 in Kalimantan and 0.917 in Brunei (Table 3). In Malaysian Borneo and Kalimantan the random forest model not

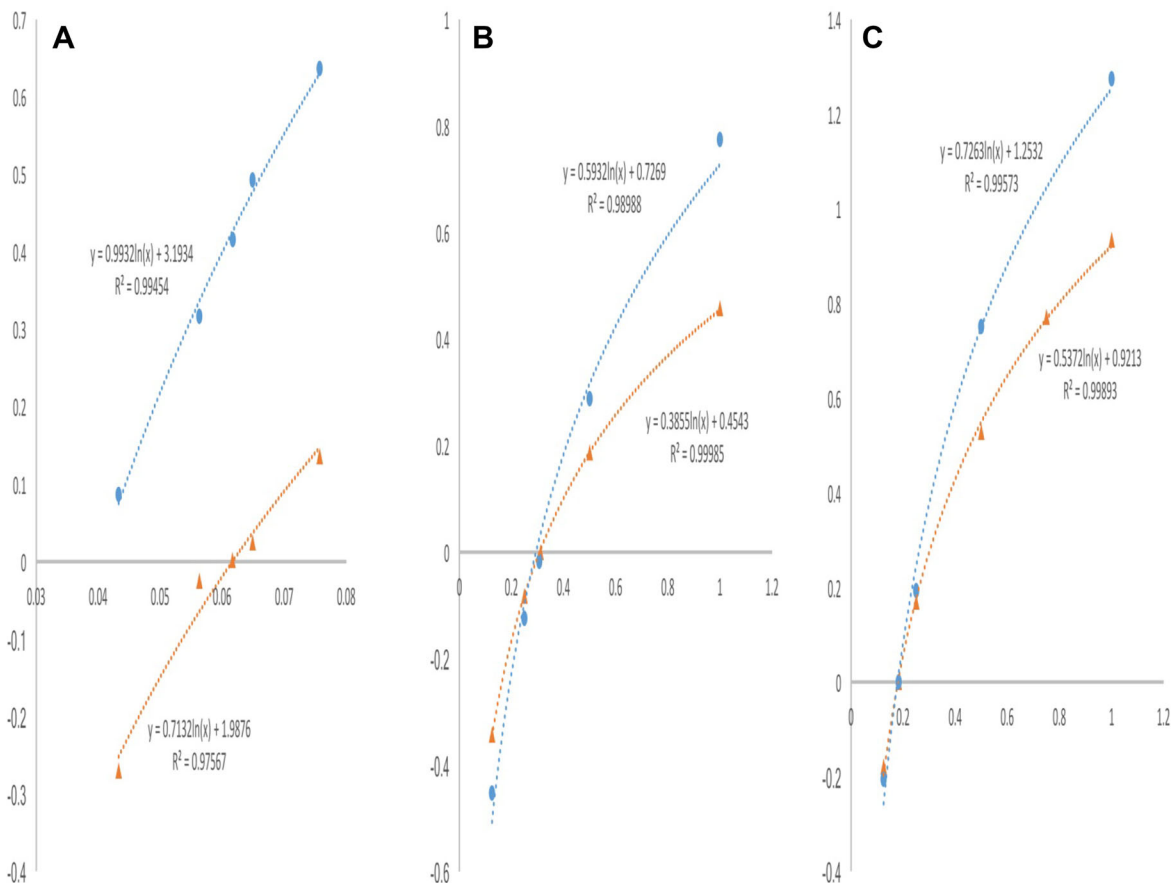


Fig. 2 Calibration of forest loss versus forest persistence points for **a** Brunei, **b** Malaysian Borneo and **c** Kalimantan. The x-axis reports the ratio of forest loss to forest persistence points in the training sample. The vertical line reflects the actual ratio in observed forest loss versus forest persistence points between 2000 and 2010. The y-axis reports the ratio of over/under

prediction of forest loss. Positive numbers reflect over prediction of forest loss by that proportion, while negative numbers reflect under prediction. The random forest calibration and trend line is shown with circular markers, with a logarithmic fit. The logistic regression calibration and trend line is shown with triangular markers, with logarithmic fit

Table 2 Tabulation of number of forested pixels persisting and lost, and proportion loss between 2000 and 2010 for the three nations in the Borneo area

	Forest pixels persisting 2000–2010	Forest pixels lost 2000–2010	Proportion loss 2000–2010	Predicted proportion loss
Brunei	15,123	812	0.051	0.059
Malaysian Borneo	334,489	104,221	0.238	0.234
Kalimantan	959,577	175,264	0.154	0.154

Predicted proportion Loss is the expected value of predicted loss obtained by averaging the probability of loss in the predicted probability map for each nation

Table 3 Model performance assessed with the TOC for assessing accuracy of continuous predicted probability surfaces, relative to the naïve model

	Logistic regression	Random forest	Random forest excluding landscape metrics	Naïve
Brunei				
maxAUC	0.866	0.927	0.823	0.845
AUC	0.852	0.917	0.805	0.833
minAUC	0.839	0.907	0.787	0.821
Malaysia				
maxAUC	0.820	0.933	0.901	0.720
AUC	0.816	0.931	0.898	0.720
minAUC	0.811	0.925	0.894	0.720
Kalimantan				
maxAUC	0.880	0.931	0.895	0.848
AUC	0.875	0.928	0.890	0.848
minAUC	0.870	0.925	0.886	0.848

Logistic Regression—calibrated logistic regression using the same variables as the calibrated random forest model; Random Forest including landscape metrics—calibrated random forest model; RF(no FRAG)—calibrated random forest model excluding landscape metrics calculated by FRAGSTATS. Naïve—naïve model expressing deforestation risk as a power function of distance from forest edge. The TOC was run with 100 equal-interval thresholds

including FRAGSTATS landscape variables was the second highest performing model with AUC of 0.898 and 0.890 respectively. In contrast, in Brunei the calibrated logistic regression and the naïve models both outperformed the random forest model not including landscape metrics based on TOC AUC (0.853 and 0.833 vs. 0.805). In the two large study areas, Kalimantan and Malaysian Borneo, the naïve model based on a power function of distance to forest from its edge performed weakest (Table 3).

The display of the TOC curves (Fig. 3) illustrate several things not illustrated by the AUC table alone. TOC shows for each threshold the hits, misses, false alarms, and correct rejections (Pointius and Si 2014). The vertical distance from the horizontal axis to the TOC curve equals Hits, while the vertical distance

from the TOC curve to the Hits + Misses line equals Misses. The maximum height of the TOC curve is equal to Hits + Misses, which indicates the size of the forest loss. The horizontal distance from the bounding parallelogram on the left to the TOC curve equals false alarms, and the horizontal distance from the TOC curve to the parallelogram on the right equals correct rejections. The most notable is the compression of the TOC curves in Kalimantan, with the naïve model performing quite close to the logistic regression model, and the logistic regression model quite close to the random forest models. This indicates that the pattern of deforestation in Kalimantan is quite similar to the naïve model as a function of distance from the edge of past deforestation. In contrast the higher spread of the curves in Malaysian Borneo indicates

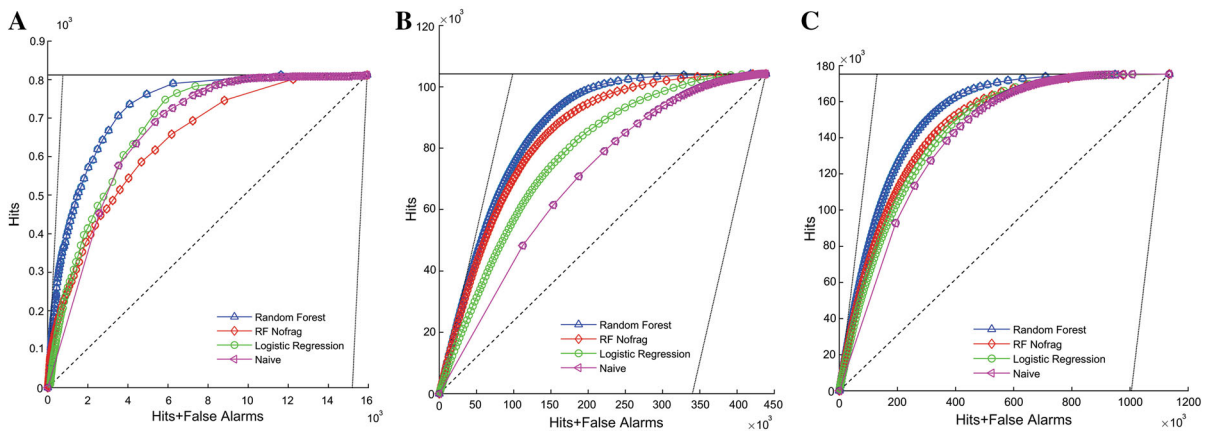


Fig. 3 TOC curves showing comparative model performance among the logistic regression, random forest, random forest without fragmentation variables, and the naïve model. Higher

that forest loss in this study area is driven by factors not reflected well in the naïve model. In addition, in Malaysian Borneo there is a relatively large gap between the logistic regression and the two random forest models as well, indicating relatively higher performance of random forest as compared to logistic regression.

Model interpretation

Our analysis produced nine different predictive models, consisting of random forest with and without landscape metrics, and logistic regression models for each of the three nations comprising Borneo. In the interest of space we briefly describe the highest performing model (random forest with landscape metrics) for each nation here, and present full information on all nine models and their interpretation in Appendix 1.

Brunei model

The ten most important variables in the calibrated random forest Brunei model based on model improvement ratio were aggregation index at 10 km radius, edge density at 10 km radius, proportion of *peat swamp forest* at 40 km radius, proportion of *water* at 50 km radius, proportion of *plantation or regrowth* at 1 km radius, proportion of *lowland forest* at 40 km radius, Shannon Diversity at 10 km radius, Topographical Roughness at 40 km radius, Elevation and proportion of *lowland mosaic* at 40 km radius

model performance is indicated by stronger convex curvature toward the upper left corner of the plot space. **a** Brunei, **b** Malaysian Boreno, **c** Kalimantan

(Figure A1). Of the ten most influential variables in the prediction of deforested vs not-deforested cells in Brunei between 2000 and 2010, as judged by model improvement ratio, five had positive monotonic relationships with increasing probability of deforestation with increasing value of the variable (edge density at 10 km radius, proportion of *peat swamp forest* at 30 km radius, proportion of *water* at 50 km radius, proportion of *plantation or regrowth* at 1 km radius, Shannon Diversity Index at 10 km radius; Figure A2). Two variables, including the most influential variable (aggregation index at 10 km radius), had negative monotonic relationships with deforestation risk, such that deforestation risk decreases as the value of these variables increase. Two variables had unimodal relationships such that the frequency of deforestation was maximum at intermediate values (proportion of *lower montane forest* at 40 km radius and Distance to Population Centre). We produced visualization of the pattern of predicted forest loss probability across Brunei (Figure A3), with zoomed-in view of two areas showing the pattern of observed loss in relation to the predicted probability of loss (Figure A4).

Malaysian Borneo model

The ten most important variables, based on model improvement ratio, in the calibrated Malaysian Borneo model were proportion of *lowland mosaic* at 20 km radius, proportion of *plantation or regrowth* at 30 km radius, edge density at 20 km radius, Elevation, proportion of *montane mosaic* at 50 km radius, patch

density at 30 km radius, proportion of *lower montane forest* at 50 km radius, Shannon diversity index at 40 km radius, focal mean population density at 100 km radius and topographical roughness at 40 km radius (Figure A6). Of the ten most influential variables in the calibrated random forest model in Malaysian Borneo, six had positive monotonic relationships wherein frequency of forest loss increased as the value of the variable increased (proportion of *lowland mosaic* at 20 km radius, proportion of *plantation or regrowth* at 30 km radius, edge density at 20 km radius, patch density at 30 km radius, Shannon diversity at 40 km radius, and focal mean population density at 100 km radius). All of these showed strongly non-linear relationships with rapid initial rise followed by asymptotic flattening (Figure A7). Four of the ten most influential variables in the calibrated Malaysian Borneo random forest model showed monotonic negative relationships where frequency of deforestation between 2000 and 2010 declined as the value of the variable increased. These mostly showed what appear to be negative exponential shapes, with rapid decline at low values of x followed by flattening as the x variable increased (elevation, proportion of *montane mosaic* at 50 km radius, proportion of *upper montane forest* at 50 km radius, and topographical roughness at a 40 km radius; Figure A7). We produced visualization of the pattern of predicted forest loss probability across Brunei (Figure A8), with zoomed-in view of two areas showing the pattern of observed loss in relation to the predicted probability of loss (Figure A9).

Kalimantan model

The ten most important variables, based on model improvement ratio, for Kalimantan were elevation, patch density at 40 km radius, proportion of *lowland mosaic* at 50 km radius, proportion of *lower montane forest* at 20 km radius, proportion of *plantation or regrowth* at 50 km radius, edge density at 10 km radius, focal mean population density at 100 km radius, topographical roughness at 50 km radius, proportion of *lowland open* at 50 km radius and Shannon's diversity index at 40 km radius (Figure A11). Of the ten most important variables in the calibrated random forest model predicting deforested versus not-deforested cells between 2000 and 2010, seven had positive monotonic relationships (patch

density at 40 km radius, proportion of *lowland mosaic* at 50 km radius, proportion of *plantation or regrowth* at 50 km radius, edge density at 10 km radius, focal mean population density at 100 km radius, proportion of *lowland open* at 50 km radius, and Shannon diversity at 40 km radius; Figure A12). As in the Malaysian Borneo model, many of these were strongly nonlinear relationships. The remaining three top variables all had negative monotonic relationships (elevation, proportion of *lower montane forest* at 20 km, topographical roughness 50 km radius). As in the Malaysian Borneo model, these were strongly nonlinear, with most showing a negative exponential shape (Figure A12). We produced visualization of the pattern of predicted forest loss probability across Kalimantan (Figure A13), with zoomed-in view of two areas showing the pattern of observed loss in relation to the predicted probability of loss (Figure A14).

Predicted forest loss risk in 2010–2020

We produced a map of predicted forest loss risk in the 2010–2020 time period (Fig. 4) and calculated the expected value of amount of forest loss and as proportion of each nation by averaging the pixel values of predicted forest loss risk across all pixels that were forest in 2010 (Table 4). The rate of forest loss predicted in the 2010–2020 period was substantially higher in Malaysian Borneo than in Kalimantan or Brunei (23.2 vs. 15.9%, Table 4). The predicted rate of forest loss in 2010–2020 was quite similar into that observed in 2000–2010 for Malaysian Borneo (23.2 vs. 23.3%) and Kalimantan (15.9 vs. 15.4%), while the predicted amount of forest loss in Brunei increased to 15.9% from the observed amount of 5.9% in 2000–2010.

We calculated the proportion of each study region in four bins of degree of risk (<25%, 25–50%, 50–75% and >75% predicted probability of forest loss in the 2010–2020 period; Fig. 5). This analysis showed a difference in the distribution of risk among study regions, which is visually apparent from inspection of Fig. 4. Specifically, Malaysian Borneo has a “flatter” distribution of proportion of pixels across the four risk bins, indicating that a larger proportion of the land area of Malaysian Borneo has moderate risk (33% of forest has 25–50% predicted probability of forest loss in 2010–2020), and a lower proportion has low risk (60%

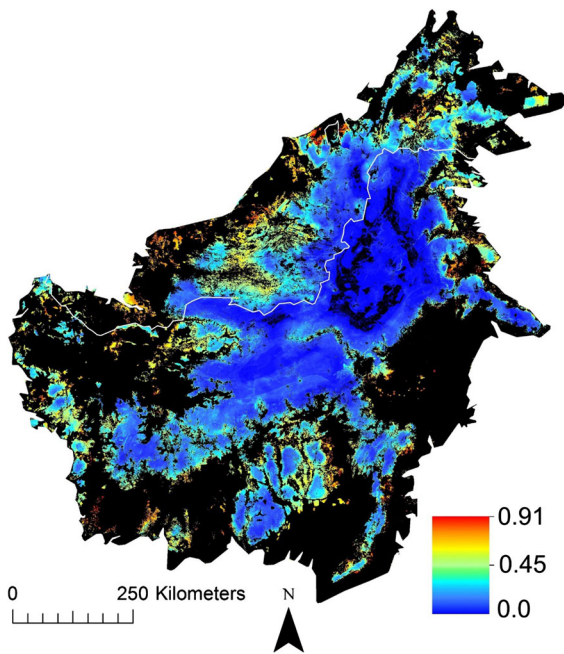


Fig. 4 Predicted probability of deforestation in the period 2010–2020, obtained by applying the calibrated random forest models developed in the 2000–2010 period to the landscape data in 2010 for each of the three study nations (Brunei, Malaysian Borneo, Kalimantan). *Black* represents areas of non-forest in 2010. (Color figure online)

Table 4 Expected number of pixels predicted to be deforested in each of the three study regions in the period between 2010 and 2020 total and as a proportion of the amount of forest existing in 2010

	Expected number pixels deforested 2010–2020	Expected proportion deforested 2010–2020
Malaysia	77601	0.232
Kalimantan	152470	0.159
Brunei	2537	0.159

has less than 25% probability of forest loss in 2010–2020), than the other two regions. In Fig. 4 this is seen as a broader area of yellow, indicating moderate risk, and less area of blue (low risk). In contrast, Kalimantan shows a pattern with a high proportion of the land area with low risk (75.5% less than 25% probability of forest loss 2010–2020) and a comparatively low proportion of area with intermediate risk (17.9% has between 25 and 50% probability of forest loss between 2010 and 2020). Both Malaysian

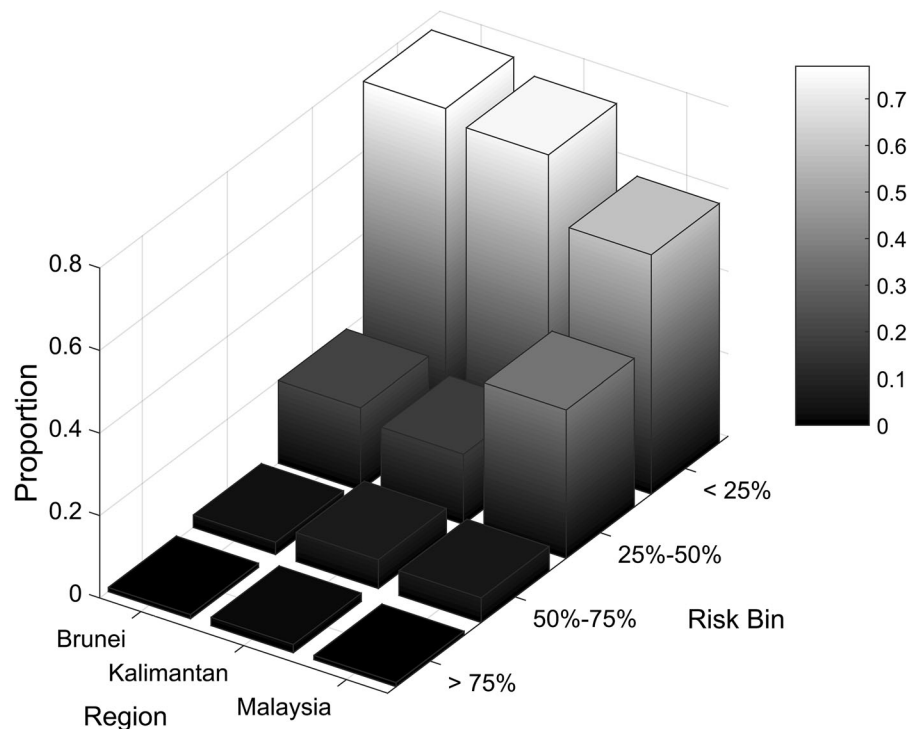
Borneo and Kalimantan have similar proportion of high risk, with 6.7 and 6.3% of forest area having between 50 and 75% probability of forest loss in 2010–2020.

Discussion

The goal of this study was to identify the multi-scale drivers of forest loss across Borneo, and then to apply this model to predict and map the risk of forest loss between 2010 and 2020. This appears to be the first study to: (1) quantitatively compare the performance of random forest and logistic regression in predicting forest loss risk, (2) conduct formal multi-scale model optimization to predict forest loss risk, (3) formally evaluate the effect of the ratio of forest loss to persistence cells on the predicted probability of forest loss, and (4) evaluate the utility of including landscape metrics quantifying the composition and configuration of a landscape mosaic on predicted risk of future forest loss.

We identified several hypotheses to help us explore these relationships. Contrary to our first hypothesis and the findings of Chawala et al. (2003) and Chen et al. (2004), we found that the probability maps produced by random forest do not under predict loss when the ratio of loss in the training sample was equal to that in the real landscape. On the contrary, our calibration results showed that both logistic regression and random forest produce unbiased predicted probability of forest loss only when the ratio of loss to persisting forest in the training sample matches that in the actual landscape. The bias increases substantially when the ratio departs from that of the actual landscape, with logistic regression substantially more sensitive than random forest to this bias. There are a number of implications of this observation that extend beyond the present study. Following Chawala et al. (2003) and Chen et al. (2004), most users of random forest have purposefully skewed the ratio of training points away from that of the actual population, intending to avoid bias (e.g., Evans and Cushman 2009). Our results clearly show this actually introduces bias and we advocate for future users of random forest to carefully select training data that are representative and proportional to the composition of actual landscape.

Fig. 5 Proportion of forest area existing in 2010 in each of the three study regions predicted to have risk of deforestation in the 2010–2020 time period of (1) less than 25%, (2) between 25 and 50%, (3) between 50 and 75%, and (4) greater than 75%



Second, following other researchers who found that random forest outperforms other methods for prediction and classification (e.g., Cushman et al. 2010; Evans et al. 2011; Rodriguez-Galiano et al. 2012; Schneider 2012), we expected that random forest would outperform logistic regression and the naïve model based on distance to forest edge. Consistent with this hypothesis, random forest was the highest performing method in all three nations comprising Borneo. This confirms the utility of random forest as a modelling tool for predicting forest loss, and we suggest that future studies use this powerful technique, with correct ratio of loss and persisting locations in the training sample.

Consistent with our third hypothesis we found that models that included landscape variables had substantially higher performance than models that excluded them. In all three nations the random forest model that included landscape variables outperformed the random forest model that excluded them. One of the central tenants of landscape ecology is that patterns influence processes across a range of scales (Turner 1989; Wiens 1989). However, to our knowledge, this paper is the first using multiple scale optimization (sensu McGarigal et al. 2016) of the effects of risk of

forest loss, and the first to utilize landscape metrics (e.g., FRAGSTATS, McGarigal et al. 2012). Our analysis shows that landscape pattern variables were important predictors, with the majority of the ten most influential variables in each nation measuring aspects of landscape composition or configuration.

In our fourth hypothesis we predicted that risk of forest loss would be related to predictor variables similarly across the three nations. Contrary to this expectation we noted substantial differences between the nations. Most notably, the pattern of forest loss risk in Kalimantan appears to be driven primarily by elevation and distance to the edge of previous forest loss, while that in Malaysian Borneo is more diffuse, less associated with proximity to previous forest loss, and more driven by landscape heterogeneity at broad scales. This likely reflects differences between the socio-economic drivers of forest loss in these two nations. We further explored this observation qualitatively by inspecting high resolution Google Earth satellite imagery for several landscapes in Malaysian Borneo and Kalimantan. We observed a general pattern of difference between Malaysian Borneo and Kalimantan, wherein forest landscapes in Malaysian Borneo, particularly those in Sarawak at lower

elevations, typically have large networks of logging roads that permeate currently unlogged forest, providing diffuse access for future harvest that matches the pattern of our prediction, while in Kalimantan we observed few such diffuse road networks and in most Kalimantan landscapes roads do not extend deeply into currently unlogged forest, leading to the “nibbling at the edge” pattern we observed and the similarity to the distance to edge naïve model.

Our fifth hypothesis proposed that forest loss risk would be related to (a) topography, with high risk at low elevations and flat areas, and low risk at higher elevations and steep terrain, (b) past forest loss in the surrounding region, (c) protected area status, and (d) human population parameters such as distance to large settlements and local population density, in that order of influence. Our results were mixed regarding these predictions. Only in Kalimantan were any of the topographical variables the most important based on Model Improvement Ratio, with elevation having the largest effect, and topographical roughness having the 7th highest effect. In Malaysian Borneo elevation was the fourth most influential variable and topographical roughness was the 10th, while in Brunei topographical roughness and elevation were the seventh and eighth most influential variables. In Malaysian Borneo the two most influential variables were extent of the local landscape covered in areas that have had previous forest loss (lowland mosaic and plantation/regrowth), while in Kalimantan these were the third and fifth most important predictors, which supports our expectation that extent of previous forest loss in a local landscape would be a strong predictor of future loss. In all three nations, however, we found a strong effect of landscape configuration metrics such as edge density (#3 Malaysia, #2 Brunei, #8 Kalimantan), patch density (#6 Malaysia, #2 Brunei, #2 Kalimantan) and aggregation index (#1 Brunei) and Shannon landscape diversity (#8 Malaysia, #7 Brunei, #10 Kalimantan). In all three nations these landscape configuration metrics, which measure the total heterogeneity and fragmentation of the landscape, were more important than either protected area status or human population density or distance to population centers. This suggests that heterogeneous landscape configurations created by past forest loss are highly predictive of future forest loss. Importantly, in none of the three nations was protected area status selected by the variable selection, suggesting that across Borneo

protected area effectiveness may be low in minimizing forest loss and/or that most protected areas are in locations of very low predicted risk based on other factors (such as high elevation areas that have high topographical roughness and high distance from past forest loss).

Predicted rates and patterns of risk of forest loss in 2020

We predicted that the rates and patterns of forest loss in the 2010–2020 period are likely to be similar to those seen in the 2000–2010 period. Forest loss in Indonesian Borneo is predicted to continue to move along fronts of contagious expansion from previously logged areas, spreading into the forest and leaving very few patches behind. Conversely, forest loss in Malaysian Borneo we predict will continue to expand more diffusely throughout the landscape, leading to much more extensive and heterogeneous patterns of forest loss. These differences reflect the differences in these nations in the extensiveness of forest road networks. We predicted very similar total amounts of forest loss in the 2010–2020 period as observed in the 2000–2010 period in both Malaysian Borneo and Kalimantan, with Malaysian Borneo continuing to have one of the highest forest loss rates reported in the world, and Kalimantan having a lower but still very high rate of forest loss by global standards.

Conclusion

Our results have several important implications. We confirm reports that random forest outperforms logistic regression for prediction of forest loss in our study areas. In doing so, and in contrast to common practice, we demonstrate that unbiased results were achieved only when the proportion of the loss class in the training sample matched the proportion in the actual landscape. We also confirm that multiple-scale modelling using landscape metrics as predictors in a random forest framework is a powerful approach to landscape change modelling. There is immense immanent risk to Borneo’s forests, with clear spatial patterns of risk related to topography and landscape structure that differ between the three nations that comprise Borneo. There are substantial differences in the spatial drivers and patterns of deforestation in

Kalimantan compared to Malaysian Borneo, where in the former forest loss is highly associated with the edge of previous loss, and is a spatially contagious “nibbling” on the remaining forest patches, while in the latter extensive road networks built along ridgetops lead to widespread, diffuse and highly fragmenting patterns of forest loss.

Acknowledgements The authors would like to thank everybody who has contributed to this manuscript. This work was funded by Grants to D.W.M. by the Robertson Foundation and the Recanati-Kaplan Foundation; E.A.M. is a Kaplan Scholar and supported by the Epply Foundation and Woodspring Trust. S.A.C. was supported by the US Forest Service, Rocky Mountain Research Station during this work. The authors would also like to thank the thesis examiners and anonymous reviewers who also contributed greatly.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aide TM, Clark ML, Grau HR, López-Carr D, Levy MA, Redo D, Bonilla-Moheno M, Riner G, Andrade-Núñez MJ, M. Muñiz (2013) Deforestation and Reforestation of Latin America and the Caribbean (2001–2010). *Biotropica* 45(2):262–271
- Aldwaik SZ, Onsted JA, Pontius RGJr (2015) Behavior-based aggregation of land categories for temporal change analysis. *Int J Appl Earth Obs Geoinformation* 35(B):229–238
- Baccini A, Goetz SJ, Walker WS, Laporte NT, Sun M, Sulla-Menashe D, Hackler J, Beck PSA, Dubayah R, Friedl MA, Samanta S, Houghton RA (2012) Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nat Clim Change* 2(3):182–185
- Błaszczynski JS (1997) Landform characterization with geographic information systems. *Photogramm Eng Remote Sens* 63(2):183–191
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Carr DL (2004) Proximate population factors and deforestation in tropical agricultural frontiers. *Popul Environ* 25(6):585–612
- Chambers CL, Cushman SA, Medina-Fitoria A, Martínez-Fonseca J, Chávez-Velásquez M (2016) Influences of scale on bat habitat relationships in a forested landscape in Nicaragua. *Landsc Ecol* 31(6):1299–1318
- Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Lavrač N, Gamberger D, Todorovski L, Blockeel H (eds) *Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, September 22–26, 2003. Berlin, Heidelberg, Springer, Berlin Heidelberg, pp 107–119
- Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data. <http://oz.berkeley.edu/users/chenchao/666.pdf>
- Chowdhury RR (2006) Driving forces of tropical deforestation: The role of remote sensing and spatial models. *Singap J Trop Geogr* 27(1):82–101
- Clark ML, Aide TM, Grau HR, Riner G (2010) A scalable approach to mapping annual land cover at 250 m using MODIS time series data: a case study in the dry chaco ecoregion of South America. *Remote Sens Environ* 114(11):2816–2832
- Cohen J (1960) A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 20(1):37–46
- Crookston NL, Finley AO (2008) yaImpute: an R package for kNN imputation. *J Stat Softw* 23(10):1–16
- Cushman SA, Evans JS, McGarigal K, Kiesecker JM (2010) Toward Gleasonian Landscape Ecology: From Communities to Species, From Patches to Pixels. Fort Collins, CO, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station
- Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification ecology. *Ecology* 88(11):2783–2792
- De’ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81(11):3178–3192
- Echeverria C, Coomes DA, Hall M, Newton AC (2008) Spatially explicit models to analyze forest loss and fragmentation between 1976 and 2020 in southern Chile. *Ecol Model* 212(3):439–449
- Evans JS, Cushman SA (2009) Gradient modeling of conifer species using random forests. *Landscape Ecol* 24(5):673–683
- Evans JS, Oakleaf J (2012) Geomorphometry & gradient metrics toolbox (ArcGIS 10.0)
- Evans JS, Murphy MA, Holden ZA, Cushman SA (2011) Modeling species distribution and change using random forest. In: Drew CA (ed) *Predictive species and habitat modeling in landscape ecology: concepts and applications*. Springer, New York
- Evans JS, Oakleaf J, Cushman SA, Theobald DM (2014) An ArcGIS toolbox for surface gradient and geomorphometric modeling, version 2.0-0
- Friedl MA, Sulla-Menashe D, Tan B, Schneider A, Ramankutty N, Sibley A, Huang X (2010) MODIS Collection 5 global land cover: algorithm refinements and characterization of new datasets. *Remote Sens Environ* 114(1):168–182
- Gaveau DLA, Kshatriya M, Sheil D, Sloan S, Molidena E, Wijaya A, Wich S, Acrenaz M, Hansen M, Broich M, Guariguata MR, Pacheco P, Potapov P, Turubanova S, Meijaard E (2013) Reconciling forest conservation and logging in Indonesian Borneo. *PLoS ONE* 8(8):e69887
- Grimm R, Behrens T, Märker M, Elsenbeer H (2008) Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using random forests analysis. *Geoderma* 146(1):102–113

- Grinand C, Rakotomalala F, Gond V, Vaudry R, Bernoux M, Vieilledent G (2013) Estimating deforestation in tropical humid and dry forests in Madagascar from 2000 to 2010 using multi-date Landsat satellite images and the random forests classifier. *Remote Sens Environ* 139:68–80
- Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A, Egorov A, Chini L, Justice CO, Townshend JR (2013) High-resolution global maps of 21st-century forest cover change. *Science* 342(6160):850–853
- Harris NL, Brown S, Hagen SC, Saatchi S, Petrova S, Salas W, Hansen MC, Potapov PV, Lutsch A (2012) Baseline map of carbon emissions from deforestation in tropical regions. *Science* 336(6088):1573–1576
- Jarvis A, Reuter HI, Nelson A, Guevara E (2006) Hole-filled seamless SRTM Data V4. International Centre for Tropical Agriculture (CIAT), Cali
- Levin SA (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73(6):1943–1967
- Liaw A, Wiener M (2002) Classification and regression by random forest. *R News* 2(3):18–22
- Maeda EE, de Almeida CM, de Carvalho Ximenes A, Formaggio AR, Shimabukuro YE, Pellikka P (2011) Dynamic modeling of forest conversion: simulation of past and future scenarios of rural activities expansion in the fringes of the Xingu National Park, Brazilian Amazon. *Int J Appl Earth Obs Geoinf* 13(3):435–446
- McGarigal K, Cushman SA, Ene E (2012) FRAGSTATS v4: spatial pattern analysis program for categorical and continuous maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. <http://www.umass.edu/landeco/research/fragstats/fragstats.html>
- McGarigal K, Wan HY, Zeller KA, Timm BC, Cushman SA (2016) Multi-scale habitat modeling: a review and outlook. *Landscape Ecol* 31(6):1161–1175
- Miettinen J, Shi C, Liew SC (2011) Deforestation rates in insular Southeast Asia between 2000 and 2010. *Glob Change Biol* 17(7):2261–2270
- Miettinen J, Shi C, Tan WJ, Liew SC (2012) 2010 land cover map of insular Southeast Asia in 250-m spatial resolution. *Remote Sens Lett* 3(1):11–20
- Mon MS, Mizoue N, Htun NZ, Kajisa T, Yoshida S (2012) Factors affecting deforestation and forest degradation in selectively logged production forest: a case study in Myanmar. *For Ecol Manag* 267:190–198
- Murphy MA, Evans JS, Storfer A (2010) Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. *Ecology* 91(1):252–261
- NASA (2010) MODIS land team surface reflectance product information. National Aeronautics and Space Administration, Washington, DC
- Pontius RG, Millones M (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int J Remote Sens* 32(15):4407–4429
- Pontius RG, Parmentier B (2014) Recommendations for using the relative operating characteristic (ROC). *Landsc Ecol* 29(3):367–382
- Pontius Jr RG, Si K (2014) The total operating characteristic to measure diagnostic ability for multiple thresholds. *Int J Geogr Inf Sci* 28(3):570–583
- Pontius RG, Walker R, Yao-Kumah R, Arima RE, Aldrich S, Caldas M, Vergara D (2007) Accuracy Assessment for a Simulation Model of Amazonian Deforestation. *Anna Assoc Am Geogr* 97(4):677–695
- Pontius RG, Boersma W, Castella J-C, Clarke K, de Nijs T, Dietzel C, Duan Z, Fotsing E, Goldstein N, Kok K, Koomen E, Lippitt CD, McConnell W, Mohd Sood A, Pijawski B, Pithadia S, Sweeney S, Trung TN, Veldkamp AT, Verburg PH (2008) Comparing the input, output, and validation maps for several models of land change. *Ann Reg Sci* 42(1):11–37
- R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Rideout AJR, Joshi NP, Viergever KM, Huxham M, Briers RA (2013) Making predictions of mangrove deforestation: a comparison of two methods in Kenya. *Glob Change Biol* 19(11):3493–3501
- Riley SJ, DeGloria SD, Elliot R (1999) A terrain ruggedness index that quantifies topographic heterogeneity. *Intermt J Sci* 5:1–4
- Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J Photogramm Remote Sens* 67:93–104
- Rosa IMD, Purves D, Souza C Jr, Ewers RM (2013) Predictive modelling of contagious deforestation in the Brazilian Amazon. *PLoS ONE* 8(10):e77231
- Rozenstein O, Karnieli A (2011) Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. *Appl Geogr* 31(2):533–544
- 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
- Selvaradjou S, Montanarella L, Spaargaren O, Dent D, Filippi N, Dominik S (2005) European Digital Archive of Soil Maps (EuDASM)—metadata of the soil maps of Asia (Luxembourg: Office of the Official Publications of the European Communities)
- Soares-Filho B, Moutinho P, Nepstad D, Anderson A, Rodrigues H, Garcia R, Dietzsch L, Merry F, Bowman M, Hissa L, Silvestrini R, Maretti C (2010) Role of Brazilian Amazon protected areas in climate change mitigation. In: Proceedings of the National Academy of Sciences
- Soares-Filho BS, Nepstad DC, Curran LM, Cerqueira GC, Garcia RA, Ramos CA, Voll E, McDonald A, Lefebvre P, Schlesinger P (2006) Modelling conservation in the Amazon basin. *Nature* 440(7083):520–523
- Soares-Filho B, Rodrigues H, Follador M (2013) A hybrid analytical-heuristic method for calibrating land-use change models. *Environ Modell Softw* 43:80–87
- Svetnik V, Liaw A, Tong C, Wang T (2004) Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T (eds) Multiple classifier systems., Lecture Notes in Computer Science/Springer, Berlin, pp 334–343

- Timm BC, McGarigal K, Cushman SA, Ganey JL (2016) Multi-scale Mexican spotted owl (*Strix occidentalis lucida*) nest/roost habitat selection in Arizona and a comparison with single-scale modeling results. *Landscape Ecol* 31(6): 1209–1225
- Thapa RB, Shimada M, Watanabe M, Motohka T, Shiraiishi T (2013) The tropical forest in south east Asia: monitoring and scenario modeling using synthetic aperture radar data. *Appl Geogr* 41:168–178
- Turner MG (1989) Landscape ecology: the effect of pattern on process. *Annu Rev Ecol Syst* 20(1): 171–197
- Venema HD, Calamai PH, Fieguth P (2005) Forest structure optimization using evolutionary programming and landscape ecology metrics. *Eur J Oper Res* 164(2):423–439
- Vieilledent G, Grinand C, Vaudry R (2013) Forecasting deforestation and carbon emissions in tropical developing countries facing demographic expansion: a case study in Madagascar. *Ecol Evol* 3(6):1702–1716
- Wahyunto, BH, Bektı H, Widiastuti F (2006) Maps of peatland distribution, Area and Carbon content in Papua, 2000–2001 Wetlands International-Indonesia Programme & Wildlife Habitat Canada (WHC). Reproduced within Interactive Atlas of Indonesia Ls Forests (CDROM). World Resources Institute, Washington, DC
- Wahyunto RS, Subagjo H (2003) Maps of area of peatland distribution and carbon content in Sumatera, 1990–2002. Wetlands Int-Indonesia Programm and Wildl Habitat Canada, Bogor
- Wahyunto RS, Subagjo H (2004) Map of peatland distribution area and carbon content in Kalimantan, 2000–2002. Wetlands International—Indonesia Programme & Wildlife Habitat Canada (WHC), Bogor
- Wiens JA (1989) Spatial scaling in ecology. *Funct Ecol* 3(4):385–397