Landscape Ecol (2009) 24:587–598
DOI 10.1007/s10980-009-9343-y

REPORT

Comparing fire spread algorithms using equivalence testing and neutral landscape models

Brian R. Miranda · Brian R. Sturtevant · Jian Yang · Eric J. Gustafson

Received: 30 April 2008 / Accepted: 7 March 2009 / Published online: 20 March 2009

Abstract We demonstrate a method to evaluate the degree to which a meta-model approximates spatial disturbance processes represented by a more detailed model across a range of landscape conditions, using neutral landscapes and equivalence testing. We illustrate this approach by comparing burn patterns produced by a relatively simple fire spread algorithm with those generated by a more detailed fire behavior model from which the simpler algorithm was derived. Equivalence testing allows objective comparisons of the output of simple and complex models, to determine if the results are significantly similar. Neutral landscape models represent a range of landscape conditions that the model may encounter, allowing evaluation of the sensitivity and behavior of the model to different landscape compositions and configurations. We first tested the model for universal applicability, then narrowed the testing to assess the practical domain of applicability. As a whole, the calibrated simple model passed the test for significant equivalence using the 25% threshold. When applied to a range of landscape conditions different from the calibration scenarios, the model failed the tests for equivalence. Although our particular model failed the tests, the neutral landscape models were helpful in determining an appropriate domain of applicability and in assessing the model sensitivity to landscape changes. Equivalence testing provides an effective method for model comparison, and coupled with neutral landscapes, our approach provides an objective way to assess the domain of applicability of a spatial model.

Keywords Equivalence · Fire model · Meta-model · Neutral landscape · Scale

B. R. Miranda (✉) · B. R. Sturtevant · E. J. Gustafson
Institute for Applied Ecosystem Studies, USDA Forest Service, Northern Research Station, 5985 Highway K, Rhinelander, WI 54501, USA
e-mail: brmiranda@fs.fed.us

J. Yang
Department of Natural Resources and Environmental Science, University of Nevada-Reno, 1000 Valley Road, MS 186, Reno, NV 89512, USA

Introduction

Landscape-scale modeling is hampered by our lack of empirical understanding of key processes at broad spatial scales (Levin 1992; Schneider 1994; Urban 2005; Wu and Hobbs 2002). One solution to this problem is to "scale-up" more detailed and finer-scale models that are more closely connected to empirical data. This meta-model approach entails generalizing the finer-scale model by extracting only those components that are useful at the broader scale (Urban et al. 1999). Yet meta-modeling can be unreliable when scaling spatial processes such as dispersal and spread because spatial processes often respond to landscape structure in a nonlinear fashion (Hargrove et al. 2000; Stauffer and Aharony 1992) due to directional interactions (Strayer et al. 2003) and spatial autocorrelation (McKenzie et al. 1996). Landscape ecologists require methods to objectively evaluate whether spatial processes are appropriately scaled within a meta-modeling framework, accurately representing the spatial components of the finer-scale model.

Objective comparison between models requires a clear definition of model similarity. Equivalence testing is used to determine whether two “treatments” are practically similar (Parkhurst 2001), and may be applied for both model validation and comparison (Robinson et al. 2005; Robinson and Froese 2004). Equivalence testing is derived from bioequivalence testing [e.g., (Berger and Hsu 1996; Wellek 2003)], which is well established in biomedical research [e.g., (Zariffa et al. 2000)]. The approach is often used in the testing of generic medicines where the requirements go beyond failing to find statistical difference in effectiveness to having to demonstrate statistical equivalence. The burden of proof in equivalence tests is reversed from traditional comparison techniques [e.g., Fisher’s significance tests (Welsh et al. 1996)], using the hypothesis of dissimilarity as the null hypothesis, meaning that rejection of the null hypothesis results in a conclusion of significant similarity. Rather than a conclusion that no significant difference is observed (sometimes interpreted as similarity), these tests allow the conclusion that there is significant similarity or equivalence. Robinson and Froese (2004) and Robinson et al. (2005) provide thorough descriptions of the differences and benefits of equivalence testing, with examples applying equivalence testing to non-spatial ecological models. Here we apply their approach to compare output from spatial models using several response variables that describe the spatial patterns.

The domain of applicability for a spatial process sensitive to landscape pattern can be determined by evaluating its behavior across a range of landscape conditions. Neutral landscape models are models that generate an expected spatial pattern in the absence of specific landscape processes (Gardner et al. 1987; With and King 1997). Neutral landscapes can objectively represent a range of conditions that a landscape model may encounter by controlling the proportions and aggregation of landscape classes, allowing a modeler to evaluate the sensitivity and behavior of a given spatial process to different elements of pattern. Here we illustrate how equivalence testing can be used in conjunction with neutral landscapes to evaluate first the universality, and second the domain of applicability of a modeled spatial process in a meta-modeling context.

Our objective was to demonstrate a method to evaluate the degree to which a simple fire spread algorithm approximated the burn patterns simulated by a more detailed fire-behavior model from which it was derived, using a combination of neutral landscapes and equivalence testing. We calibrated the algorithm using a single landscape condition and range of weather conditions, and then evaluated the calibrated algorithm across a range of landscapes and weather conditions using a two stage process—first testing for universality of the algorithm and then by evaluating its domain of applicability if it failed the universality test. Our approach has practical application for meta-modeling strategies that simplify spatial processes to scale-up fine-scaled behavior to broader spatial scales.

Methods

Meta-modeling approach

Fire modeling varies in complexity from simple statistical models to complex, physically-based fire behavior models (Albright and Meisner 1999). Our research evaluating landscape change in response to fire disturbance regimes over long time periods requires simulated burn patterns that are sensitive to landscape patterns of fuel types, but where the details of fire behavior within a burn event are not relevant and merely add significant computation time. We therefore calibrated a simple fire spread algorithm, outlined below, to match the shape characteristics of fires produced by a more detailed fire behavior model [FARSITE; (Finney 2004)]. We extracted the model components relating specifically to spread, and generalized them to the level of detail suitable for our simulation model. We then evaluated the generality of the calibrated model by applying both models to neutral landscapes with different proportions and spatial arrangements of fuel types and compared the results using equivalence tests.

Model descriptions

FARSITE is a deterministic model (when spotting is not enabled) that simulates the spread and behavior of fires in response to terrain, fuel, and weather conditions. Fire spread is simulated based on physical equations for spread rates (Albini 1976), crown fire spread (Van Wagner 1993), spotting (Albini 1979), point-source fire acceleration (Forestry Canada Fire Danger Group 1992), and fuel moisture (Hartford and Rothermel 1991). In our case we were focused on modeling surface fire regimes, so we did not enable spotting in our FARSITE simulations.

Our simple spread algorithm is an adaptation of the minimum travel time method proposed by Finney (2002) that uses directional wind bias and maximum (i.e., downwind) spread rates defined by wind strength and fuel type to produce a relative time cost surface from the ignition point to each cell in a potentially burned zone. Wind bias is defined by the ratio of the major to minor axis of an ellipse that increases as a function of five classes of wind strength. Wind bias normal to the fire front and relative to the ignition point is then estimated for each wind class using Eqs. 3 and 5 of Finney (2002). This wind bias is converted into an index (WIND) ranging between 0 and 1 by dividing all values by the maximum wind bias (i.e., directly downwind from the ignition point). Fuel types are also assigned to five classes, and the user assigns the maximum rate of spread ($\mathrm{RATE}_{\mathrm{MAX}}$) to each fuel type and wind speed combination. In this implementation, the effects of slope on spread rate and elliptical dimensions are not considered.

For a given fire event, an ignition cell is identified, a wind speed and direction are randomly selected from a user-defined distribution, and a fire size is randomly selected from a lognormal distribution (Yang et al. 2008). Wind bias relative to the ignition point (WIND) is estimated for each cell that could potentially burn given the wind speed and fire size selected. $\mathrm{RATE}_{\mathrm{MAX}}$ is then assigned to each cell based on the fuel class for that cell and the wind speed for the event. Actual spread rate ($\mathrm{RATE}_{\mathrm{ACT}}$) is calculated by:

$$\mathrm{RATE}_{\mathrm{ACT}} = \mathrm{WIND}^{W_{\mathrm{W}}} \times \mathrm{RATE}_{\mathrm{MAX}}^{W_{\mathrm{F}}} \tag{1}$$

where W_{W} and W_{F} are calibration parameters that determine the relative importance of wind bias and fuel type, respectively, in generating the fire shape. The inverse of $\mathrm{RATE}_{\mathrm{ACT}}$ is used as a cost surface to calculate the minimum travel time to each pixel from the ignition location. Minimum travel time is then used to clip the fire extent to the preselected fire size, defining its final shape.

Model parameterization

Our application landscape was the Lakewood subdistrict (780 km^2) of the Chequamegon-Nicolet National Forest (CNNF) in Wisconsin, USA (Sturtevant et al. 2009). This landscape contains a mixture of cover types including deciduous, coniferous, and mixed forests as well as open field and wetlands that vary substantially in their relative flammability (Sturtevant and Cleland 2007). Each type was assigned to a standard fuel model (Albini 1976) based on the expert opinion of local fire management officers. These fuel types were then stratified into the five fuel classes based on their relative rates of spread (Table 1; Fig. 1). Wind speed data from a local

Table 1 Landscape proportions for standard fuel models (Albini 1976) and the five aggregated fuel classes used in the simple fire algorithm, for the original landscape, and the [L + 8] and [L + 1] landscapes

Standard fuel model	Fuel class	Mean spread rate (m/min)	Original landscape	[L + 8] landscape	[L + 1] landscape
0 (No fuel)	0	0	0.0326	0.0249	0.0286
1 (Short grass)	4	37.6	0.1883	0.1436	0.2883
4 (Chaparral)	5	41.4	0.0130	0.0099	0.0114
8 (Closed timber litter)	1	0.5	0.5788	0.6788	0.5075
9 (Hardwood litter)	2	2.0	0.1759	0.1341	0.1542
10 (Timber with litter & understory)	2	2.0	0.0106	0.0081	0.0093
11 (Light logging slash)	3	2.8	0.0008	0.0006	0.0007

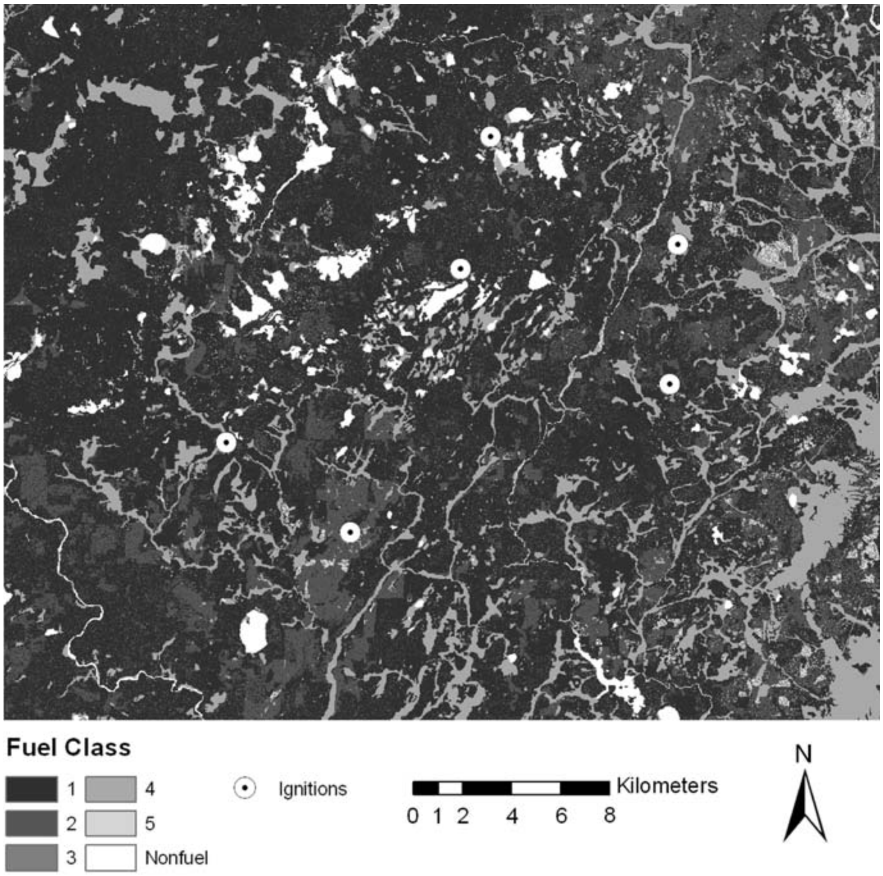

Fig. 1 Map of the fuel classes and ignition locations in the calibration landscape

weather station were used to assign midpoint wind speeds to the five wind strength classes according to percentiles defined by the US Forest Service fire danger rating system, where low, moderate, high, very high, and extreme wind speeds correspond with 50, 75, 90, 95, and 98th percentiles, respectively. RATE_{MAX} parameters were estimated by simulating fires on uniform fuels in FARSITE and averaging the downwind spread rates for each fuel and wind class for the range of possible fuel moisture conditions for the application landscape. The rate estimates could also be calculated using BehavePlus (Andrews et al. 2005). Wind bias for each wind strength class was estimated using Eq. 79 of the Canadian fire behavior prediction system (Forestry Canada Fire Danger Group 1992).

Model calibration

Our simple fire spread algorithm is computationally efficient, but differs from Finney (2002) and FARSITE because maximum wind bias is always directly downwind from the ignition point, rather than locally estimated from a dynamic fire front. This difference results in shape differences for fires responding to barriers and other fuel heterogeneity. To minimize these differences, we calibrated our spread algorithm by running the model over a range of combinations (0.50–1.00 by 0.05 for each variable) of the relative weights of wind and fuel factors (W_{W} and W_{F}) affecting fire spread (Eq. 1), with the goal of making fire patch metrics statistically equivalent across a range of fire weather scenarios (see "Statistical methods").

The fire algorithm was calibrated across nine different fire weather scenarios representing two factors (wind speed and fuel moisture) each with three different levels (low, high, and extreme) based on local fire danger ratings. Six independent fire events were simulated per scenario, where two ignition points were located on each of the three most common fuel types, always in the same locations across scenarios. The ignition locations were randomly located within the fuel types, with a restriction preventing ignitions to be located within 5,000 m of the map edge to minimize the occurrence of fires burning to the edge of the map. Fires were simulated in FARSITE using a 12-h simulated burn period, and burned areas were calculated for each simulated fire event. Fire events were then simulated using our fire spread algorithm for the same ignition points using identical wind strength and direction, and stopped spreading when the size reached the FARSITE size. The stochastic nature and broad scale of the intended application of this spread algorithm within the forest landscape succession fire models [e.g., LANDIS simulation model (He and Mladenoff 1999)] make the actual cell-by-cell agreement of the burned areas less important than the spatial characteristics of the burn patches. As long as the algorithm produces realistic patterns of burn patches in response to fuel configuration and composition on a landscape, it is not critical that the actual same sites are burned. Therefore, rather than evaluating the overlap of the burned areas between the two models, we chose five variables to compare fire events simulated by each model: the proportion of burned area in each of the three most common fuel classes, and the shape complexity and elongation of the simulated fire patches, estimated using SHAPE and CIRCLE metrics from FRAGSTATS (McGarigal and Marks 1995), respectively.

Model evaluation

After completing the calibration of model parameters using the real landscape, we then assessed the domain of applicability of the calibrated spread algorithm within novel landscapes by applying it to neutral landscapes and testing the equivalence of the burn patterns. The first stage of this evaluation was to determine if the model was universally applicable across a broad range of landscape conditions. If the model performed equally well (i.e., statistically equivalent) across the full range of conditions tested, we would conclude that the model was universally applicable across those conditions and the evaluation would end with this stage. If the model failed to perform equally well, the second stage was designed to evaluate the practical domain of applicability of the model, defined as the range of conditions for which the model outputs were statistically equivalent to FARSITE outputs for the same conditions. In this second stage we tested a narrower range of conditions starting with those most similar to the calibrated condition, where the model would presumably perform the best. By systematically changing the conditions from the calibrated condition, the bounds of the domain of applicability were identified.

For the first stage of evaluation (universal applicability), we initially generated five neutral landscapes, representing three levels of fuel type proportions and three different fuel configurations, using the program RULE (Gardner 1999). RULE uses a midpoint displacement algorithm (Saupe 1988) to generate multi-fractal landscapes, with the degree of aggregation controlled by a user-defined parameter, H (Gardner 1999). One random (R) and two multi-fractal (F) maps were generated with fuel types and proportions identical to the Lakewood landscape, but with different levels of spatial aggregation: random (noted as R0), $H = 0.3$ (F3), and $H = 0.6$ (F6) (Table 2). Two additional multi-fractal maps were created with moderate spatial aggregation (F3) but with modified fuel proportions representing either a 10% increase of the most common and least flammable fuel type (Fuel Model 8, noted as [L + 8]) or a 10% increase in the second most common and most flammable fuel type (Fuel Model 1, [L + 1]) (Tables 1, 2). For each evaluation landscape, six combinations of wind strength and fuel moisture levels (low, high or extreme) were used to define the evaluation scenarios, with six

Table 2 Landscape composition and configuration combinations used for model evaluation

		Fuel model composition		
		Original (same as real landscape)	Fuel Model 8 + 10% [L + 8]	Fuel Model 1 + 10% [L + 1]
Fuel configuration (aggregation) index (H)	0	R0	N/A	N/A
	0.1	F1	F1 [L + 8]	F1 [L + 1]
	0.2	F2	F2 [L + 8]	F2 [L + 1]
	0.3	F3	F3 [L + 8]	F3 [L + 1]
	0.6	F6	N/A	N/A

"N/A" indicates combinations that were not evaluated. "Fuel Model 8 + 10%" and "Fuel Model 1 + 10%" indicate the proportion of Fuel Model 8 or Fuel Model 1 has been increased by 0.10 from the original fuel composition (Table 1) independent fires simulated for each scenario, resulting in 36 fires simulated for each landscape. For our goal of initially assessing the universal applicability of the model, including all nine wind and fuel moisture combinations and altered proportions for the R0 and F6 landscapes, was not necessary and would have more than doubled the simulations required. Using this sample of scenarios allowed us to evaluate the full range of conditions, and to use the output data to make relative comparisons.

For the second stage of evaluation (domain of applicability), we generated six additional neutral landscapes representing two additional levels of spatial aggregation: $H = 0.1$ (F1) and $H = 0.2$ (F2), with the same three levels of fuel proportions (Original, [L + 1], [L + 8]) used for the F3 landscapes (Table 2). In combination with the F3 landscapes, these landscapes provided a narrower range of conditions similar to the calibration condition, varying in fuel configuration and proportion, with which to assess a more refined domain of applicability for the model.

Statistical methods

We applied a regression-based test for equivalence (Robinson et al. 2005) to test for similarity between each fire metric for our simple algorithm (observed) and those generated by FARSITE (predicted). This approach tests both the intercept and slope of a regression between the observed and predicted values. This regression framework uses the intercept to evaluate population-level agreement (unbiasedness) and the slope to evaluate point-to-point agreement (i.e., the individual pairs of observations are similar). Because our data did not meet assumptions of normality, we used the non-parametric bootstrap method described by Robinson et al. (2005) with 10,000 replicate estimates of the intercept and slope. Following the suggestion of Robinson et al. (2005), we report the minimum interval of equivalence (MIE) as the smallest interval that would lead to rejection of the null hypothesis of dissimilarity. MIE provides a measure of how close the test was to rejecting the null hypothesis—the equivalent of reporting confidence intervals for other statistical tests—with a standardized value. Units of MIE are proportions of the mean [e.g., 0.25 indicates the interval of equivalence is mean ± (0.25 × mean)]. We conducted equivalence tests for the five variables in two stages. We first calculated MIE values for the proportions of burned area in each of the three most common fuel classes and for the SHAPE and CIRCLE indices. To give SHAPE, CIRCLE, and fuel proportions equal weight in our evaluations, and because each MIE value had the same units, we averaged the MIE values for the three fuel classes, giving each of the three main outputs (SHAPE, CIRCLE, average Fuel) one value for comparison. With a target joint size of 0.05, the Bonferroni-adjusted α for each of the three main tests (SHAPE, CIRCLE, average Fuel) was $\alpha = 0.017$. The Bonferroni-adjusted α for each of the three individual fuel tests was $\alpha = 0.006$. Each of these α values was further adjusted within the R program to account for the two regression tests (slope and intercept). All equivalence testing was performed using code for R (R Development Core Team 2007) provided by Andrew Robinson (package Equivalence v. 0.4.1; http://www.biometrics.mtu.edu/CRAN/web/packages/equivalence/index.html).

Although we evaluated both unbiasedness and point-to-point agreement in the equivalence tests, we focus here on the point-to-point agreement because this assesses whether the population-level agreement is for the right reason. We also found point-to-point agreement tests to consistently have larger MIE values than the unbiasedness tests, implying that point-to-point agreement was essentially the limiting factor for determining equivalence. To have one final MIE value to test, we calculated an overall MIE value as the average of the SHAPE, CIRCLE, and average Fuel MIE values for the point-to-point agreement test. Robinson et al. (2005) use intervals of equivalence of ±25% in their example equivalence tests, which we used as a threshold for determining statistical equivalence.

Results

Calibration

During calibration, we found the lowest average MIE of 0.20 for point-to-point agreement when W_W had a weight of 1.00 and W_F had a weight of 0.65 (Table 3). In the calibrated model, the CIRCLE variable had the largest MIE for point-to-point agreement, while the average fuel proportion had the best agreement. As a

Table 3 Minimum intervals of equivalence (MIE) for unbiasedness (intercept) and point-to-point agreement (slope) for each variable for the calibration using the real landscape, and universal evaluation with all neutral landscapes, with the calibrated relative weights for wind and fuel of 1.00 and 0.65, respectively

Variable	Calibration (real landscape)		Universal test (neutral landscapes)	
	Unbiasedness	Point-to-point	Unbiasedness	Point-to-point
n	54	54	288	288
SHAPE	**0.07**	**0.16**	**0.15**	0.47
CIRCLE	**0.05**	0.37	**0.09**	0.55
Fuel proportion average	**0.06**	**0.08**	**0.09**	**0.14**
Average	**0.06**	**0.20**	**0.11**	0.39

Units of MIE are proportions of the mean [e.g., a value of 0.25 indicates the interval of equivalence is mean ± (0.25 × mean)]. Numbers in bold mark values at or below the 0.25 threshold, indicating significant equivalence

whole, the calibrated model (average MIE) passed the test for significant equivalence for both unbiasedness and point-to-point agreement using the ±25% threshold. Of the individual components, only the CIRCLE variable failed the point-to-point test. In addition to quantitative similarities, the burned areas appeared visually similar (Fig. 2).

Evaluation for universal applicability

Evaluation scenarios on neutral landscapes show that model performance decreased when the conditions changed from the calibration scenarios (Table 3). Unbiasedness and point-to-point agreement decreased for each of the variables, despite the increase in sample size. As in the calibration scenarios, the CIRCLE index had the poorest point-to-point agreement and the fuel proportions had the best. Only the fuel proportions passed the test for equivalence, while the SHAPE, CIRCLE and the model as a whole failed this test.

By breaking down the evaluation scenarios into groupings by landscape, we identified conditions under which the model performed best and worst. Examining

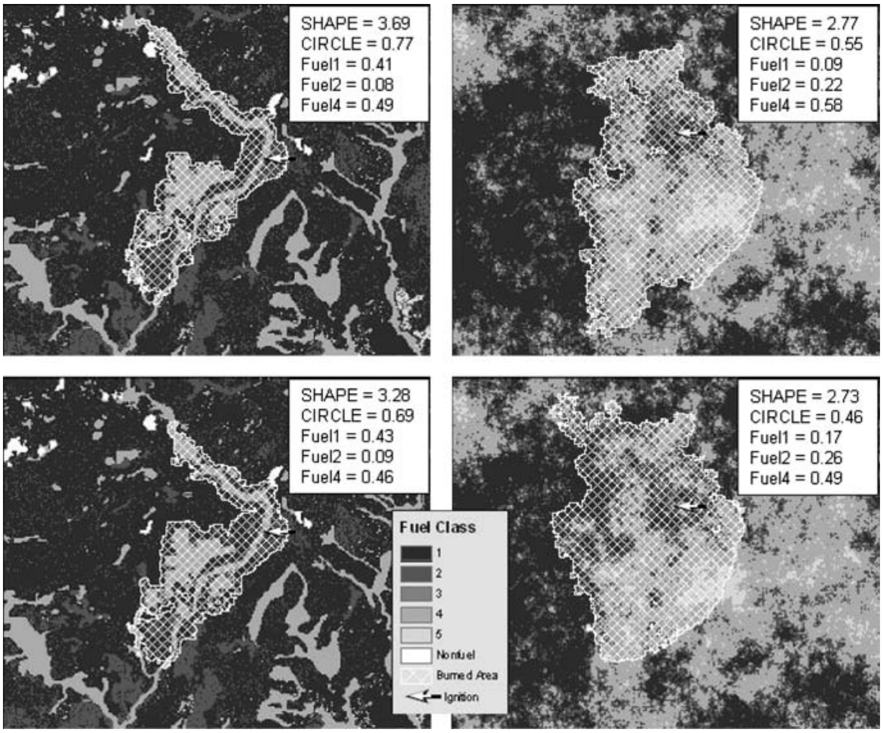

Fig. 2 Example fire burn patches for the calibration landscape (*left*) and F2 evaluation landscape (*right*), generated from FARSITE (*top*) and the simple model (*bottom*). The ignition location and wind direction is indicated by the *arrow*. The patch attributes are included for each burned area. The SHAPE and CIRCLE indices come from FRAGSTATS, and Fuel 1, Fuel 2, and Fuel 4 indicate the proportion of the burned area in Fuel Classes 1, 2, and 4, respectively

the three landscapes with the original fuel proportions that differed only in fuel configuration, we observed that the model as a whole performed considerably better in the F3 landscape than either then R0 or F6 landscapes (Table 4). Examining the three F3 landscapes that differed only in fuel proportions, we observed that the model performed better on both of the alternate proportions than on the original proportion landscape, and showed the best results for the [L + 8] landscape (Table 5). All of the five landscapes we evaluated failed the overall (average) tests for equivalence for point-to-point agreement at the ±25% threshold, although some individual tests for the fuel proportions were below the threshold.

Evaluation for domain of applicability

After concluding that the model was not universally acceptable, we narrowed our focus to finding the range of conditions, if any, where the model outputs would pass the overall equivalence tests. We narrowed the range of fuel configurations being tested to those most similar to the calibration condition (F1, F2, F3), and included the same three fuel proportions. None of the combinations of fuel configuration and fuel proportions that we evaluated passed the test for point-to-point agreement at the ±25% threshold (Fig. 3). The F2 landscape had consistently better agreement than the F1 and F3 landscapes, and the model performed best on the [L + 8] landscapes. For individual fires, burned areas with quantitatively similar attributes did not necessarily look similar visually (Fig. 2). In all evaluation landscapes, the average MIE for point-to-point agreement was larger than the MIE for the unbiasedness test (Tables 4, 5). With only one exception each, the point-to-point MIE for the CIRCLE variable was larger than for the SHAPE and fuel proportions for each landscape.

Table 4 Minimum intervals of equivalence (MIE) for unbiasedness (intercept) and point-to-point agreement (slope) for each variable for the separate evaluation landscapes with the original fuel proportions and varied configurations, with the calibrated relative weights for wind and fuel of 1.00 and 0.65, respectively

Variable	R0		F3 (original)		F6	
	Unbiasedness	Point-to-point	Unbiasedness	Point-to-point	Unbiasedness	Point-to-point
n	36	36	36	36	36	36
SHAPE	**0.23**	0.75	**0.17**	0.54	**0.19**	0.62
CIRCLE	**0.15**	1.00	**0.13**	0.79	**0.15**	0.94
Fuel proportion average	**0.13**	1.05	**0.16**	**0.15**	0.26	**0.25**
Average	**0.17**	0.93	**0.15**	0.49	**0.20**	0.60

Units of MIE are proportions of the mean [e.g., a value of 0.25 indicates the interval of equivalence is mean ± (0.25 × mean)]. Numbers in bold mark values at or below the 0.25 threshold, indicating significant equivalence

Table 5 Minimum intervals of equivalence (MIE) for unbiasedness (intercept) and point-to-point agreement (slope) for each variable for the separate evaluation landscapes with the F3 configuration and varied fuel proportions, with the calibrated relative weights for wind and fuel of 1.00 and 0.65, respectively

Variable	F3 [L + 8]		F3 [L + 1]	
	Unbiasedness	Point-to-point	Unbiasedness	Point-to-point
n	36	36	36	36
SHAPE	**0.15**	0.42	**0.24**	0.74
CIRCLE	**0.12**	0.61	**0.21**	0.54
Fuel proportion average	**0.11**	**0.11**	**0.14**	**0.09**
Average	**0.13**	0.38	**0.20**	0.46

See Table 4 for results for the original fuel proportions. Units of MIE are proportions of the mean [e.g., a value of 0.25 indicates the interval of equivalence is mean ± (0.25 × mean)]. Numbers in bold mark values at or below the 0.25 threshold, indicating significant equivalence

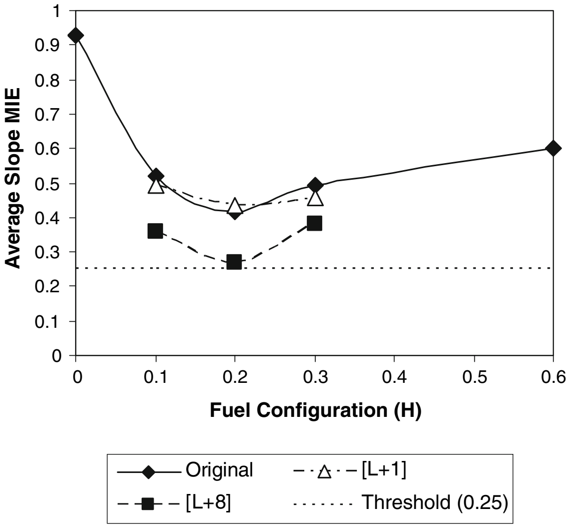

Fig. 3 Average minimum intervals of equivalence (*MIE*) for point-to-point agreement tests for landscapes across three levels of fuel proportions (original, [L + 8], [L + 1]) and three levels of fuel configuration (F1, F2, F3), with the calibrated relative weights for wind and fuel of 1.00 and 0.65, respectively. Units of MIE are proportions of the mean [e.g., a value of 0.25 indicates the interval of equivalence is mean ± (0.25 × mean)]. The *dotted line* represents the maximum threshold for significant equivalence (0.25)

Discussion

The equivalence tests in our two-stage process answer three important questions. How well can we calibrate the model? Is the model sensitive to landscape changes? What is the domain of applicability of the model?

Calibration

Equivalence tests provided an objective means for selecting the best parameter values during calibration. Averaged MIE values (average of SHAPE, CIRCLE, and average Fuel) provided a single, globally-relevant number to guide the calibration. Alternatively we could have based the calibration on MIE for a single variable, rather than averaging MIE values across multiple variables, if one specific variable were more critical than others. Our calibrated model performed reasonably well on the target landscape, with the average MIE values for both unbiasedness and point-to-point agreement passing the ±25% equivalence test. The CIRCLE variable alone failed the equivalence test, suggesting the model as a whole would still be useful for applications where accurately representing patch elongation was not critical. This calibration approach identified the parameter values with the greatest average equivalence. Any individual fire patch may be different between the model outputs, but as a whole, the fires patterns were as equivalent as possible under the tested conditions.

Evaluation

Although it would be ideal for our simplified model to be significantly equivalent to the complex model in all aspects of the output, the nature of scaling up a model (in this case, simplifying the fire spread procedure) implies that there will be differences. Whether or not the resulting differences are acceptable is the decision of the model user, and this process provides objective information to help make that decision. Our evaluation showed that this simplified model is sensitive to changes in landscape structure and composition. In fact, our model failed to produce significantly equivalent results for any landscape other than the calibration landscape. This implies that the meta-model should not be used for simulations on landscapes that differ much from the calibrated condition, without first recalibrating the model parameters.

Our evaluation of the domain of applicability failed to identify any neutral landscapes where the model would produce statistically equivalent results. However, this evaluation can still provide valuable information by looking at the relative changes in average MIE as the landscapes change. Evaluation showed that the model equivalence within the F2 landscape was most similar to that observed in the calibration (i.e., real) landscape. The relative sensitivity of burn patterns to fuel composition and fuel configuration observed for the F2 landscape can indicate how sensitive fire patterns would be to analogous changes within the real landscape. For example, a 50% change in fuel connectivity ($H = 0.2$ versus $H = 0.1$ or 0.3) results in changes up to 24% in average MIE (Fig. 3). Meanwhile, a 10% change in fuel proportions [L + 8] can change average MIE by 35%. We can infer from these results that the model can be more sensitive to changes in fuel proportions than configurations within that domain.

The choice of landscapes used for both calibration and evaluation are important. Preliminary calibration trials (not shown) applied to our neutral landscapes instead of the real landscape, resulted in different calibrated values for W_W and W_F. This difference reinforces the sensitivity of this particular model to the landscape configuration of fuels, and also suggests that the range of neutral landscapes considered depart substantially from the real landscape. Li et al. (2004) determined that neutral landscapes may not accurately represent some aspects of real landscapes, though other methods for neutral landscape generation are available [e.g., (Gardner and Urban 2007)]. Despite some of the limitations of neutral landscapes, the ability to represent multiple landscapes and systematically alter the landscape properties provides a strong foundation for equivalence testing of alternative models of spatial processes as they interact with spatial structure inherent within landscapes.

While our evaluation explicitly evaluated the domain of applicability of our simplified model with respect to fuel type proportions and configuration, other dimensions of fire spread were not evaluated—though they could be using a similar approach. For example, our simplified model did not consider either topographic influence on fire spread or spotting behavior often associated with crown fire behavior. As such it should be conservatively applied to situations where topography is not a strong influence on fire behavior, and where fire dynamics are constrained to surface fires. Additional equivalence tests applied outside of this conservative domain, using the approach outlined here, could address the range of appropriate applicability in these other dimensions.

Advantages and disadvantages

Our approach provides some advantages over alternative methods for defining the relative value and applicability of a model. One alternative method in the field of fire modeling is to compare simulated burn patterns with patterns of real fires (Fujioka 2002). One advantage of the model to model comparison approach we used is that the models can be tested across a range of fuel and weather conditions, and is not limited to testing conditions of a limited number of real fire events. The model to model comparison, however, relies on the assumption that the model used as the reference (in our case FARSITE) is “correct”. By calibrating the meta-model to the output of the more complex model, the assumptions of the complex model are effectively carried into the meta-model.

Equivalence testing provides a more robust assessment of whether the model outputs are practically and statistically equivalent, rather than the traditional approach of failing to find them to be different (Robinson and Froese 2004). We have outlined how the results of multiple variables can be combined into one measure (average MIE) of how similarly the models are performing. Using spatial pattern indices for comparison variables, rather than simply assessing overlap of the burn patches, allowed us to focus on the pattern attributes most important to our applications. Our output had examples that had similar spatial attributes, but would not be deemed similar based on a simple measure of their overlap (Fig. 2).

Neutral landscape models can represent a range of landscape conditions that a spatial model may encounter. These landscapes provide a similar advantage as the model to model comparison, in that they remove the restrictions associated with working with a single real landscape. Neutral landscapes combined with equivalence testing provide an objective way to assess the domain of applicability of a model.

The approach we have presented here also has some statistical shortcomings that could be improved. For example, an overall test of equivalence incorporating multiple response variables (the equivalent of MANOVA with a reversed hypothesis) would be superior to averaging MIE values for multiple variables. Such multivariate equivalence tests have been developed in biomedical research (Hauck et al. 1995), but as far as we know have not been applied in an ecological context. Another advantage that could be adopted from the bioequivalence literature is allowing multiple variables to have different equivalence thresholds, referred to as multiple endpoints (Pocock et al. 1987). Although informal evaluations showed that our results were not highly sensitive to sample size, differences in sample size among independent factors can make direct comparisons of test results difficult. The bootstrap method of Robinson et al. (2005) could be modified to address this issue by randomly subsampling data so that test results by factor are based on the same sample size.

Acknowledgments We thank Andrew Robinson for preliminary advice and thoughtful comments on an earlier draft, in addition to providing his Equivalence package for R. Comments and suggestions from two anonymous reviewers led to a substantially improved manuscript. We thank John Stanovick for his valuable consultations. This research was supported by funding from the National Fire Plan.

References

Albini FA (1976) Estimating wildfire behavior and effects. USDA Forest Service Intermountain Forest and Range Experiment Station, General Technical Report GTR-INT-30

Albini FA (1979) Spot fire distance from burning trees—a predictive model. USDA Forest Service, General Technical Report INT-56

Albright D, Meisner BN (1999) Classification of fire simulation systems. Fire Manage Notes 59(2):5–12

Andrews PL, Bevins CD, Seli RC (2005) BehavePlus fire modeling system, version 3.0: user's guide revised. USDA Forest Service Rocky Mountain Research Station, RMRS-GTR-106WWW

Berger RL, Hsu JC (1996) Bioequivalence trials, intersection-union tests, and equivalence confidence sets. Stat Sci 11(4):283–319. doi:10.1214/ss/1032280304

Finney MA (2002) Fire growth using minimum travel time methods. Can J For Res 32:1420–1424. doi:10.1139/x02-068

Finney MA (2004) FARSITE: fire area simulator—model development and evaluation. USDA Forest Service, Rocky Mountain Research Station, Research Paper RMRS-RP-4 revised

Forestry Canada Fire Danger Group (1992) Development and structure of the Canadian Forest Fire Behavior Prediction System. Forestry Canada, Science and Sustainable Development Directorate, Information Report ST-X-3

Fujioka FM (2002) A new method for the analysis of fire spread modeling errors. Int J Wildland Fire 11(3–4):193–203. doi:10.1071/WF02004

Gardner RH (1999) RULE: a program for the generation of random maps and the analysis of spatial patterns. In: Klopatek JM, Gardner RH (eds) Landscape ecological analysis: issues and applications. Springer, New York, pp 280–303

Gardner RH, Urban DL (2007) Neutral models for testing landscape hypotheses. Landscape Ecol 22:15–29. doi:10.1007/s10980-006-9011-4

Gardner RH, Milne BT, Turner MG, O'Neill RV (1987) Neutral models for the analysis of broad-scale landscape pattern. Landscape Ecol 1(1):19–28. doi:10.1007/BF02275262

Hargrove WW, Gardner RH, Turner MG, Romme WH, Despain DG (2000) Simulating fire patterns in heterogeneous landscapes. Ecol Modell 135(2–3):243–263. doi:10.1016/S0304-3800(00)00368-9

Hartford RA, Rothermel RC (1991) Moisture measurements in the Yellowstone Fires in 1988. USDA Forest Service, Research Note INT-396

Hauck WW, Hyslop T, Anderson S, Bois FY, Tozer TN (1995) Statistical and regulatory considerations for multiple measures in bioequivalence testing. Clin Res Regul Aff 12(4):249–265. doi:10.3109/10601339509019618

He HS, Mladenoff DJ (1999) Spatially explicit and stochastic simulation of forest-landscape fire disturbance and succession. Ecology 80(1):81–99

Levin SA (1992) The problem of pattern and scale in ecology. Ecology 73:1943–1967. doi:10.2307/1941447

Li X, He HS, Wang X, Bu R, Hu Y, Chang Y (2004) Evaluating the effectiveness of neutral landscape models to represent a real landscape. Landsc Urban Plan 69:137–148. doi:10.1016/j.landurbplan.2003.10.037

McGarigal K, Marks BJ (1995) FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. USDA Forest Service Pacific Northwest Research Station, General Technical Report PNW-GTR-351

McKenzie D, Peterson DL, Alvarado E (1996) Extrapolation problems in modeling fire effects at large spatial scales: a review. Int J Wildland Fire 6(4):165–176. doi:10.1071/WF9960165

Parkhurst DF (2001) Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. Bioscience 51:1051–1057. doi:10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2

Pocock SJ, Geller NL, Tsiatis AA (1987) The analysis of multiple endpoints in clinical trials. Biometrics 43(3):487–498. doi:10.2307/2531989

R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Robinson AP, Froese RE (2004) Model validation using equivalence tests. Ecol Modell 176:349–358. doi:10.1016/j.ecolmodel.2004.01.013

Robinson AP, Duursma RA, Marshall JD (2005) A regression-based equivalence test for model validation: shifting the burden of proof. Tree Physiol 25:903–913

Saupe D (1988) Algorithms for random fractals. In: Peitgen H-O, Saupe D (eds) The science of fractal images. Springer, New York, pp 71–113

Schneider DC (1994) Quantitative ecology: spatial and temporal scaling. Academic Press, San Diego

Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor and Francis, Washington DC

Strayer DL, Ewing HA, Bigelow S (2003) What kind of spatial and temporal details are required in models of heterogeneous systems? Oikos 102(3):654–662. doi:10.1034/j.1600-0706.2003.12184.x

Sturtevant BR, Cleland DT (2007) Human and biophysical factors influencing modern fire disturbance in northern Wisconsin. Int J Wildland Fire 16(4):398–413. doi:10.1071/WF06023

Sturtevant BR, Miranda BR, Yang J, He HS, Gustafson EJ, Scheller RM (2009) Studying fire mitigation strategies in multi-ownership landscapes: balancing the management of fire dependent ecosystems and fire risk. Ecosystems. doi:10.1007/s10021-009-9234-8

Urban DL (2005) Modeling ecological processes across scales. Ecology 86(8):1996–2006. doi:10.1890/04-0918

Urban DL, Acevedo MF, Garman SL (1999) Scaling fine-scale processes to large-scale patterns using models derived from models: meta-models. In: Mladenoff DJ, Baker WL (eds) Spatial modeling of forest landscape change. Cambridge University Press, Cambridge, pp 70–98

Van Wagner CE (1993) Prediction of crown fire behavior in two stands of jack pine. Can J For Res 23:442–449. doi:10.1139/x93-062

Wellek S (2003) Testing statistical hypotheses of equivalence. Chapman & Hall/CRC Press, New York

Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol Modell 88:297–308. doi:10.1016/0304-3800(95)00113-1

With KA, King AW (1997) The use and misuse of neutral landscape models in ecology. Oikos 79(2):219–229. doi:10.2307/3546007

Wu J, Hobbs R (2002) Key issues and research priorities in landscape ecology: an idiosyncratic synthesis. Landscape Ecol 17:355–365. doi:10.1023/A:1020561630963

Yang J, He HS, Sturtevant BR, Miranda BR, Gustafson EJ (2008) Comparing effects of fire modeling methods on simulated fire patterns and succession: a case study in the Missouri Ozarks. Can J For Res 38:1290–1302. doi:10.1139/X07-235

Zariffa NMD, Patterson SD, Boyle D, Hyneck M (2000) Case studies, practical issues and observations on population and individual bioequivalence. Stat Med 19:2811–2820. doi:10.1002/1097-0258(20001030)19:20<2811::AID-SIM547>3.0.CO;2-P