Chapter 8
Lessons and Challenges in Land Change Modeling Derived from Synthesis of Cross-Case Comparisons

Robert Gilmore Pontius Jr., Jean-Christophe Castella, Ton de Nijs, Zengqiang Duan, Eric Fotsing, Noah Goldstein, Kasper Kok, Eric Koomen, Christopher D. Lippitt, William McConnell, Alias Mohd Sood, Bryan Pijanowski, Peter Verburg, and A. Tom Veldkamp

Abstract This chapter presents the lessons and challenges in land change modeling that emerged from years of reflection and numerous panel discussions at scientific conferences concerning a collaborative cross-case comparison in which the authors have participated. We summarize the lessons as nine challenges grouped under three themes: mapping, modeling, and learning. The mapping challenges are: to prepare data appropriately, to select relevant resolutions, and to differentiate types of land change. The modeling challenges are: to separate calibration from validation, to predict small amounts of change, and to interpret the influence of quantity error. The learning challenges are: to use appropriate map comparison measurements, to learn about land change processes, and to collaborate openly. To quantify the pattern validation of predictions of change, we recommend that modelers report as a percentage

R.G. Pontius Jr. (✉)
Clark University, Worcester, MA, USA

Michigan State University, East Lansing, MI, USA

Universiti Putra Malaysia, Serdang, Malaysia

Purdue University, West Lafayette, IN, USA

University of Twente, Enschede, The Netherlands
e-mail: rpontius@clarku.edu

J.-C. Castella
Institut de Recherche pour le Développement, Montpellier, France
e-mail: j.castella@ird.fr

T. de Nijs
National Institute for Public Health and the Environment, De Bilt, The Netherlands
e-mail: Ton.de.Nijs@rivm.nl

Z. Duan
China Agricultural University, Beijing, China
e-mail: Duanzengqiang@aliyun.com

of the spatial extent the following measurements: misses, hits, wrong hits and false alarms. The chapter explains why the lessons and challenges are essential for the future research agenda concerning land change modeling.

Keywords CLUE • CLUE-S • Environment Explorer • Geomod • Land Transformation Model • Land change • Land Use Scanner • LUCC • Map • Model • Prediction • SAMBA • SLEUTH • Validation

8.1   Introduction

The first author of this chapter extended an open invitation to the community of land change modelers to participate in a cross-case comparison of spatially explicit land change modeling applications. The focus was the assessment of pattern validation

E. Fotsing
Computer Science, University of Dschang, Bandjoun, Cameroon
e-mail: fosting@gmail.com

N. Goldstein
Lawrence Livermore National Laboratory, Livermore, CA, USA
e-mail: goldstein@navigant.com

K. Kok
Soil Geography and Landscape, Wageningen University, Wageningen, The Netherlands
e-mail: Kasper.Kok@wur.nl

E. Koomen • P. Verburg
Spatial Analysis & Modelling, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Vrije Universiteit Amsterdam, Institute for Environmental Studies, Amsterdam, The Netherlands
e-mail: ekoomen@feweb.vu.nl; Peter.Verburg@ivm.vu.nl

C.D. Lippitt
Geography & Environmental Studies, The University of New Mexico, Albuquerque, NM, USA
e-mail: clippitt@unm.edu

W. McConnell
Center for Systems Integration and Sustainability, Michigan State University,
East Lansing, MI, USA
e-mail: mcconn64@msu.edu

A. Mohd Sood
Universiti Putra Malaysia, Selangor, Malaysia
e-mail: ms_alias@putra.upm.edu.my

B. Pijanowski
Forestry and Natural Resources, Purdue University, West Lafayette, IN, USA
e-mail: bpijanow@purdue.edu

A.T. Veldkamp
Faculty of Geo-Information Science and Earth Observation, University of Twente,
Enschede, The Netherlands
e-mail: a.veldkamp@twente.nl

of the mapped output of such models, so the invitation requested that participants submit for any case study three maps of land categories: (1) a reference map of an initial time 1 that a land change model used for calibration, (2) a reference map of a subsequent time 2 that could be used for validation, and (3) a prediction map of same time 2 that the land change model produced. Ultimately, we compiled 13 cases from nine countries, which were submitted from seven different laboratories. Pontius et al. (2008) derived and applied metrics to compare those various cases. We presented our work at several scientific conferences. Pontius et al. (2008) has been cited more than 396 times as of September 2017 according to scholar.google.com, thus has had a substantial influence on the constantly growing field of land change modeling (Paegelow et al. 2013). A frequent initial reaction that audiences have when they first hear about our exercise is to ask "Which model is best?" However, the exercise never intended to rank the models. The audience's unintended reaction has been one of the inspirations for this follow-up chapter. The popularity of the question indicates that we must be careful to interpret the results properly, because the purpose of the exercise can be easily misinterpreted. We have found that the exercise's methods and results inspire quite disparate conclusions from various scientists. The purpose of the exercise was to gain insight into the scientific process of modeling, in order to learn the most from our modeling efforts. Therefore, this chapter shares the lessons that survived after years of reflection on both participation in the cross-case comparison and interactions with colleagues.

Figure 8.1 shows how we think of the lessons in terms of the flows and feedbacks of information among the various components of modeling. The figure begins with the landscape in the upper left corner. Scientists create data to summarize the landscape. There is a tremendous amount of information that scientists can derive from simply analyzing the maps from two or more time points (Aldwaik and Pontius 2013; Runfola and Pontius 2013). Scientists anticipate that they can learn even more by engaging in a modeling procedure that produces a dynamic simulation of land change. Scientists usually use a conceptual understanding of landscape dynamics to guide the selection or production of algorithms that express those dynamics. This chapter uses the word "model" to refer to such a set of algorithms, and the word "case" to refer to an application of the model to a particular study site. One way to assess a case is to examine the output that the model produces. Ultimately, a major purpose of the analysis is for scientists to learn from the measurements of the data and the outputs from the model. Scientists can use this learning to revise the mapping, the modeling and/or the measurements of the data and the model's output. The components of Fig. 8.1 reflect the structure of this chapter in that this chapter's Methods section summarizes the techniques to measure both the data and the model's output, while the subsequent Results and Discussion section presents the most important lessons, organized under the themes of mapping, modeling, and learning.
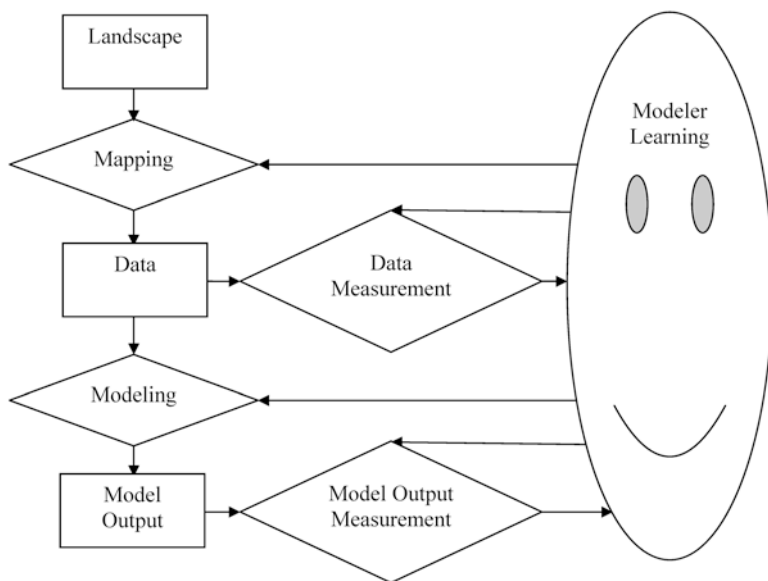
Fig. 8.1 Conceputal diagram to illustrate flows and feedbacks of information among components and procedures for a systematic analysis. *Rectangles* are components of the research system; *diamonds* are procedures; the *oval* is the modeler whose learning can inform methods of mapping, modeling and measuring

8.2 Methods

All of the models have been published in peer-reviewed journals and books. Raster maps have been submitted by scientists from the laboratories that developed the models. Collectively, the sample of models and their applications cover a range of some of the most common modeling techniques such as statistical regression, cellular automata, and machine learning. SAMBA is the single agent-based model in the collection. Table 8.1 offers specific characteristics of the nine models used for the 13 cases. These cases offer illustrations of these models that have been applied with various objectives, extents and resolutions. The model characteristics in Table 8.1 are necessary for proper interpretation. Geomod, Logistic Regression, and Land Transformation Model (LTM) use maps for which each pixel shows the land as either undeveloped or developed. These three models predict a single transition from the undeveloped category to the developed category. The other six models use maps of more than two categories to predict multiple transitions. For seven of the models, the user can set exogenously the quantity of each land cover category for the predicted map, and then the model predicts the spatial allocation of the land categories. SLEUTH and SAMBA do not have this characteristic. The cases that derive from LTM, CLUE-S, and CLUE use the quantity of each category in the reference map of time 2 as input to the model. For these cases, the model is assured to simulate the correct quantity of each category at time 2, thus the purpose of the

Table 8.1 Characteristics of the nine models as applied in the 13 cases

Model	Predicted transitions	Exogenous quantity	Used year 2 quantity	Pure pixels	Case	Literature
Geomod	Single	Yes	No	Yes	Worcester	Pontius et al. (2001), Pontius and Malanson (2005), Pontius and Neeti (2010)
SLEUTH	Multiple	No	No	Yes	Santa Barbara	Dietzel and Clarke (2004), Goldstein (2004), Silva and Clarke (2002)
Land Use Scanner	Multiple	Yes	No	No	Holland(8)	Hilferink and Rietveld (1999), Hoymann, (2010), Koomen and Borsboom-van Beurden (2011), Kuhlman et al. (2013)
Environment Explorer	Multiple	Optional	No	Optional	Holland(15)	de Nijs et al. (2004), Engelen et al. (2003), Verburg et al. (2004)
Logistic regression	Single	Yes	No	Yes	Perinet	McConnell et al. (2004)
SAMBA	Multiple	No	No	Yes	Cho Don	Boissau and Castella (2003), Castella et al. (2005a, b)
Land Transformation Model	Single	Yes	Yes	Yes	Twin Cities, Detroit	Pijanowski et al. (2000, 2002, 2005)
CLUE-S	Multiple	Yes	Yes	Yes	Kuala Lumpur, Haidian, Maroua	Duan et al. (2004), Fotsing et al. (2013), Tan et al. (2005), Verburg and Veldkamp (2004), Verburg et al. (2002)
CLUE	Multiple	Yes	Yes	No	Costa Rica, Honduras	de Koning et al. (1999), Kok and Veldkamp (2001), Kok et al. (2001), Veldkamp and Fresco (1996), Verburg et al. (1999)

modeling application is to predict the spatial allocation of change. Most of the models are designed to use pixels that are categorized as exactly one category, while Land Use Scanner, Environment Explorer and CLUE can use heterogeneous mixed pixels for both input and output.

Both Land Use Scanner and Environment Explorer are applied to the entire country of The Netherlands. One substantial difference between these two cases is that the number of categories in the output map for the application of Land Use Scanner is eight, while the number of categories for the application of Environment Explorer is 15. LTM, CLUE-S, and CLUE are applied to more than one study area, which allows us to see variation in how a single model can behave in various case studies. Our sample does not include cases of how a single model can produce various outputs for a single extent depending on how the model is parameterized. The possible variation due to parameterization of a single model is one reason why we do not rank the performance of the models.

Figure 8.2 shows the mapped results for each of the 13 cases. Each map in Fig. 8.2 derives from an overlay of the three maps that a modeler submitted. The first 11 of the 13 cases share the same legend, while Costa Rica and Honduras share a different legend because those two cases have mixed pixels. We encourage the profession to use the following short names for the categories in the legend of Fig. 8.2 (Brown et al. 2013). Misses are erroneous pixels due to observed change predicted as persistence. Hits are correct pixels due to observed change predicted as change. Wrong hits are erroneous pixels due to observed change predicted as change to the wrong gaining category. False alarms are erroneous pixels due to observed persistence predicted as change. Correct rejections are correct pixels due to observed persistence predicted as persistence.

Figure 8.3 summarizes the results where a segmented bar quantifies each case in terms of the legend of Fig. 8.2. Each bar is a Venn diagram where one set is the observed change and the other set is the predicted change, as the brackets illustrate for the case of Perinet. The "figure of merit" is a summary measurement that is a ratio, where the numerator is the number of hits and the denominator is the sum of hits, wrong hits, misses and false alarms (Pontius et al. 2007, 2011). If the model's prediction were perfect, then there would be perfect intersection between the observed change and the predicted change, in which case the figure of merit would be 100%. If there were no intersection between the observed change and the predicted change, then the figure of merit would be zero. Figure 8.3 orders the cases in terms of the figure of merit, which is expressed as a percent at the right of each bar. It is also helpful to consider a null model for each case. A null model is a prediction of complete persistence, i.e. no change between time 1 and time 2 (Pontius et al. 2004a). Consequently the accuracy of the null model is 100% minus the percent of observed change. Figure 8.3 shows that the accuracy of the land change model exceeds the accuracy of its corresponding null model for 7 of the 13 cases at the resolution of the raw data.

Figure 8.4 plots for each case the figure of merit versus the percentage of observed change. Figure 8.4 reveals two clusters. The tight cluster near the origin shows that all of the cases that have a figure of merit less than 15% also have an observed change less than 10%. We analyzed many factors that we suspected might explain the predictive power for the 13 cases and found that the percentage of change observed in the reference maps had the strongest relationship with predictive accuracy.
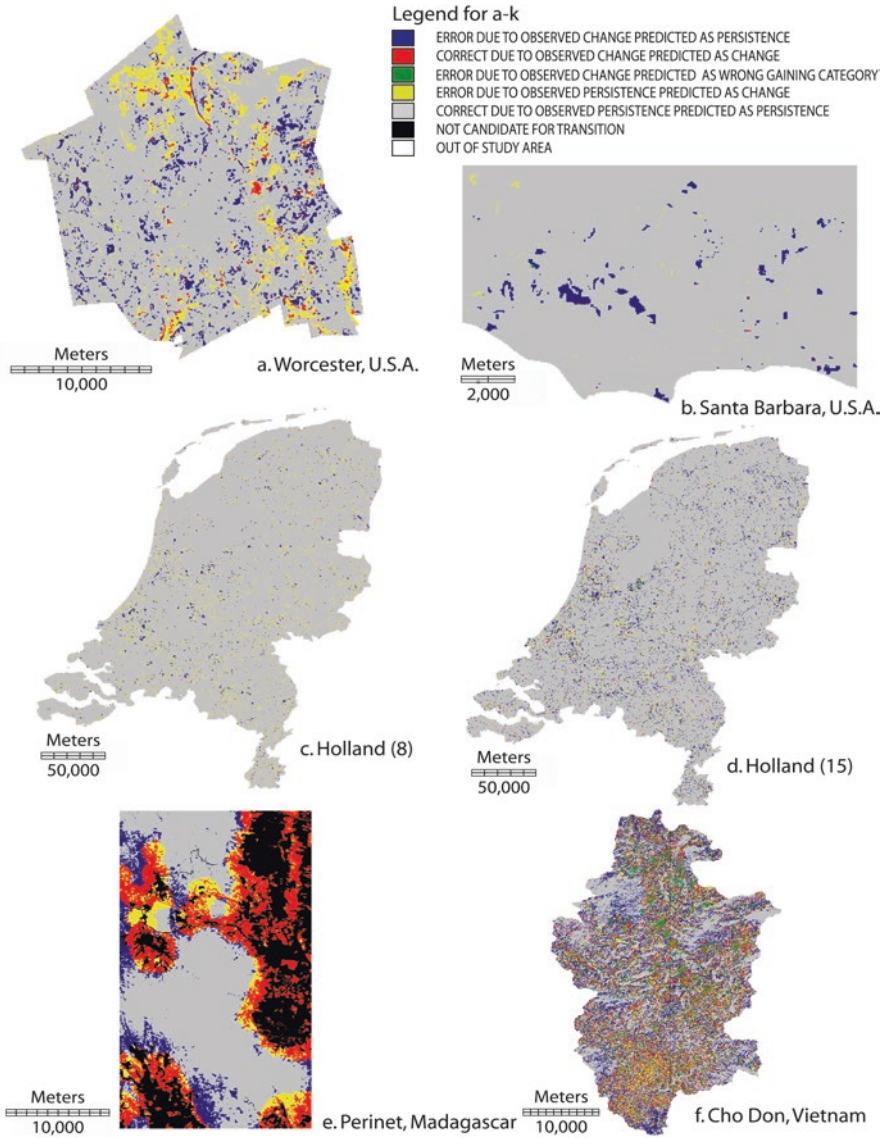
Legend for a-k

- ERROR DUE TO OBSERVED CHANGE PREDICTED AS PERSISTENCE
- CORRECT DUE TO OBSERVED CHANGE PREDICTED AS CHANGE
- ERROR DUE TO OBSERVED CHANGE PREDICTED AS WRONG GAINING CATEGORY
- ERROR DUE TO OBSERVED PERSISTENCE PREDICTED AS CHANGE
- CORRECT DUE TO OBSERVED PERSISTENCE PREDICTED AS PERSISTENCE
- NOT CANDIDATE FOR TRANSITION
- OUT OF STUDY AREA

Fig. 8.2 Maps of misses, hits, wrong hits, false alarms and correct rejections
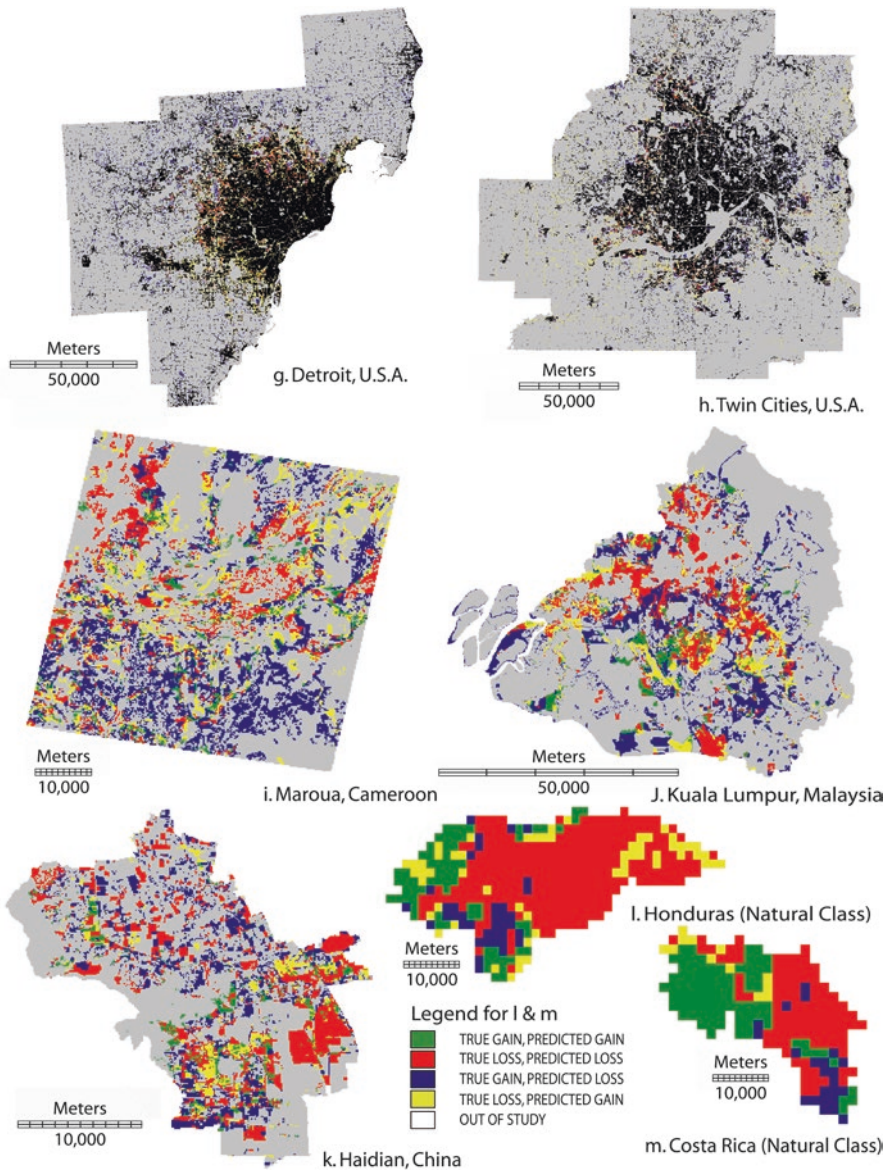
Fig. 8.2 (continued)

We have been soliciting feedback on our exercise since the initial invitation to participate in 2004. We have presented our work at five international scientific conferences: the 2004 Workshop on the Integrated Assessment of the Land System in Amsterdam The Netherlands, the 2005 Open Meeting of the Human Dimensions of Global Environmental Change Research Community in Bonn Germany, the 2006 Meeting of the Association of American Geographers in Chicago USA, the 2007 World Congress of the International Association for Landscape Ecology in
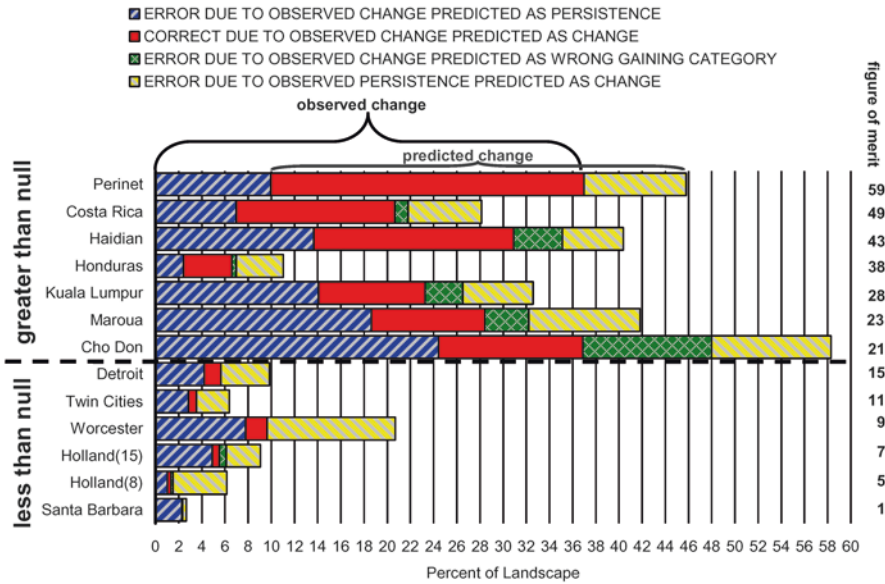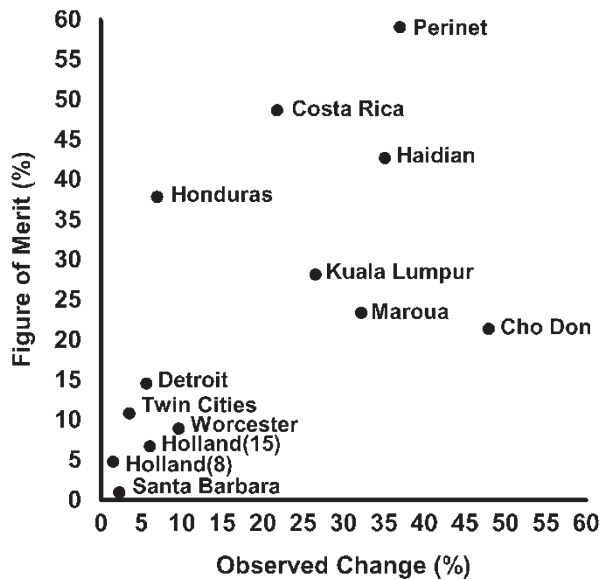
Fig. 8.3 Misses, hits, wrong hits, and false alarms for pattern validation of 13 cases. Correct rejections are 100% minus the length of the entire segmented bar. Each *bar* is a Venn diagram where the union of hits and wrong hits is the intersection of observed change and predicted change

Fig. 8.4 Relationship between predictive accuracy and observed change

Wageningen The Netherlands, and the 2007 Transatlantic Land Use Conference in Washington DC USA. There were panel discussions in Amsterdam, Chicago and Wageningen, where authors shared their experiences and audience members shared their reactions. The next section of this chapter synthesizes the lessons that have withstood more than a decade of examination of this cross-case comparison.

8.3   Results and Discussion

This section offers nine lessons. Each lesson has implications concerning the agenda for future research; therefore each lesson corresponds to a sub-section that articulates a challenge for future modeling efforts. The lessons are grouped under three themes: mapping, modeling, and learning. These groupings emerged as the authors reflected on the various types of lessons. The first theme demonstrates that the selection of the spatial extent and the production of the data have a substantial influence on the results, so scientists must pay as much attention to the mapping procedure as they do to the modeling procedure. This message reinforces known fundamental concepts in mapping, which scientists must keep at the front of their minds. The second theme concerns the modeling process. The challenges under this second theme derive from insights that have emerged specifically as a result of this cross-case exercise. They have implications for how scientists design and assess modeling procedures. The third theme focuses on learning, thus it emphasizes careful reflection on mapping and modeling. If mapping and modeling are not interpreted properly, then modelers can exert a tremendous amount of time and energy without learning efficiently. This third theme contains ideas for how modelers can maximize learning from mapping and modeling.

8.3.1   Mapping Challenges

8.3.1.1   To Prepare Data Appropriately

The decisions concerning how to format the data are some of the most influential decisions that scientists make. In some cases, scientists adopt the existing format of the available data, while in other cases scientists purposely format the data for the particular research project. Scientists must think carefully about the purpose of the modeling exercise when determining the format of the data. Formatting decisions concern the spatial, temporal and categorical scales in terms of both extent and resolution. The apparent complexity of a landscape is a function of how scientists choose to envision it, which is reflected in their mapping procedures. If scientists choose a great level of detail, then any landscape can appear to be greatly complex; while if scientists choose less detail, then the same landscape can appear simpler than what the more detailed data portray. For example, the Dutch landscape is not inherently

more complex than the Perinet landscape. However the data for Perinet were formatted to show a one-way transition from forest to non-forest while the data for Holland(15) were formatted to show multiple transitions among 15 categories based on the data formatting decisions of the modelers. One could have attempted to analyze the Dutch landscape as two categories of built versus non-built, and could have attempted to analyze the Perinet data as numerous categories of various types of uses and covers. For example, Laney (2002) chose to analyze land change in Madagascar at a much finer level of detail and deeper level of complexity than McConnell et al. (2004). Anyone can choose a great level of detail for the data that will overwhelm the computational and predictive ability of any particular model. More detail does not necessarily lead to a more appropriate case study, just as less detail does not necessarily lead to a more appropriate case study. Scientists face the challenge to select a spatial resolution, spatial extent, temporal resolution, temporal extent, and set of categories for which a model can illuminate issues that are relevant for the particular purpose of the inquiry.

Decisions concerning the format and detail of the data are fundamental for understanding and evaluating the performance of the model (Dietzel and Clarke 2004). The Holland(8) case demonstrates this clearly as it relates to the reformatting from maps that describe many heterogeneous categories within each pixel to maps that describe the single dominant category within each pixel. The Land Use Scanner model was run for heterogeneous pixels of 36 categories, and then the output was reformatted to homogenous pixels of eight categories for the three-map comparison presented in Fig. 8.2. This reformatting is common to facilitate the visualization of such mixed pixel data. A major drawback of this reformatting is that it can introduce substantial overrepresentation of categories that tend to cover less than the entire pixel but more than any other category within the pixel (Loonen and Koomen 2009). Consequently, the reformatting can also introduce substantial underrepresentation of minority categories. These artifacts due to reformatting can generate more differences between the maps than the differences that the model generates by its predicted change. Such biases substantially influenced the analysis of the Holland(8) case and caused the apparent error of quantity for the predicted change to be larger than the error of quantity for the null model.

Decisions concerning how to format the data are influential, but scientists lack clear guidelines concerning how to make such decisions. It makes sense to simplify the data to the level that the calibration procedure and validation procedure can detect a meaningful signal of land change. It also makes sense to simplify the data so that the computer algorithms focus on only the important transitions among categories, where importance is related to the practical purpose of the modeling exercise. Scientists who attempt to analyze all transitions among a large number of categories face substantial challenges. For the Santa Barbara, Holland(8), and Holland(15) cases, each particular transition from one category to another category in the reference maps occurs on less than 1% of the spatial extent. Each of these individual transitions would need to have an extremely strong relationship with the independent variables in order for a model to predict them accurately. Scientists can alleviate the challenge by aggregation from a set of numerous detailed categories to

a set of fewer coarser categories. Aldwaik et al. (2014) offer an algorithm for how to aggregate categories while maintaining the signals of land change.

Decisions concerning the data are related closely to decisions concerning the level of complexity of the models. Models that simulate only a one-way transition from one category to one other category can be simpler than models that simulate all possible transitions among multiple categories. If scientists choose to analyze very detailed data, then they may be tempted or forced to use very complex models. It is not clear whether it is worthwhile to include great detail in the data and/or in the models, because it is not clear whether more detail leads to better information or to more error.

Modelers should consider the certainty of the data, because much of the apparent land change between two time points could be due to error in the reference maps at the two time points (Enaruvbe and Pontius 2015; Pontius and Lippitt 2006; Pontius and Petrova 2010). Participating scientists suspect that error accounts for a substantial amount of the observed difference between the two reference maps for Maroua, Kuala Lumpur, and Holland(15). Scientists should use data for which there is more variation over time due to the dynamics of the landscape than due to map error. This can be quite a challenge in situations where map producers are satisfied with 85% accuracy, which implies up to 15% error, while many data sets show less than 15% land change.

8.3.1.2   To Select Relevant Spatial Resolutions

Spatial resolution is a component of data format that warrants special attention because: (1) spatial resolution can have a particularly strong influence on results, (2) spatial resolution is something that modelers usually can influence, and (3) it is not obvious how to select an appropriate spatial resolution. The spatial resolution at which landscapes are modeled is often determined by data availability and computational capacity. For example, if a satellite image dictates the resolution and extent, as it did in the Maroua case (Fotsing et al. 2013), then the boundaries of the study area and the apparent unit of analysis are determined in part by the satellite imaging system, not necessarily by the theoretical or policy imperatives of the modeling exercise. Kok et al. (2001) argue that the selection of resolution should take into consideration the purpose of the modeling application and the scales of the land change processes. For example, the Worcester case uses 30-m resolution data, but we know of no stakeholders in Worcester who need a prediction of land change to be accurate to within 30 m. Some stakeholders would like to know generally what an extrapolation of recent trends would imply over the next decade to within a few kilometers, which is a resolution at which Geomod predicts better than a null model as revealed by a multiple-resolution analysis of the model's output. Therefore, it is helpful from the standpoint of model performance to measure the accuracy of the prediction at resolutions coarser than the resolution of the raw data. Pontius et al. (2008) show that 12 of the 13 case studies have more error than correctly predicted change at the fine resolution of the raw data. However, for 7 of the 13 cases, most of the errors are due to inaccurate spatial allocation over relatively small distances. Multiple-resolution analysis shows that the errors shrink when the results are assessed at a resolution of 64 times the length of the side of the original pixels.

Errors of spatial allocation shrink as resolution becomes coarser, but errors of quantity are independent of resolution when assessed using an appropriate multiple-resolution method of map comparison (Pontius et al. 2004a).

If there is more allocation error than correctly predicted change at the resolution of the raw data, then it means that the data have a resolution that is finer than the ability of the model to predict allocation correctly. This can be a desirable characteristic because it means that the modeling exercise is not limited by the coarseness of the spatial resolution of the data. If there is more correctly predicted change than allocation error than at the resolution of the raw data, then it might be an undesirable characteristic because it might mean that the modeling exercise is limited by the coarseness of the spatial resolution of the data. The size of the error is larger than the size of correctly predicted change for 12 of 13 of our case studies at the spatial resolution of the raw data. Some scientists might conclude that the models are not accurate, while it may be more appropriate to conclude the data are more detailed than necessary.

Advances in mapping technology have made it increasingly easy to find data that have a resolution finer than is necessary to address various research questions. If data are available at the meter resolution, then it does not imply that scientists are obligated to simulate changes accurately to within a meter. It might be desirable to run the model at a fine resolution, but to analyze the output at coarser resolutions in order to find a spatial resolution for which the model predicts sufficiently given the goals of the modeling exercise.

8.3.1.3 To Differentiate Types of Land Change

Scientists should select the types of land change that are of interest before deciding which model to use, because some types of land change present particular challenges for models. It is useful to think of two major types of change: quantity difference and allocation difference. Quantity difference refers to the difference in the size of the categories in the reference maps of time 1 and time 2, while allocation difference refers to the difference in the spatial allocation of the categories given the quantity difference (Pontius et al. 2004b; Pontius and Millones 2011; Pontius and Santacruz 2014). Allocation difference exists when a category experiences loss at some places and gain at other places during a time interval. The reference maps for Holland(15), Cho Don, Haidian, Honduras and Costa Rica demonstrate more allocation than quantity difference. In particular, Costa Rica demonstrates about ten times more allocation than quantity difference. When there is substantial allocation difference in the observed data, the model is faced with the challenge to predict simultaneous gains in some pixels and losses in other pixels for a single category in order to predict the change accurately. This can be much more challenging than to predict a one-way transition from one category to one other category. For example, the Worcester, Perinet, Detroit, and Twin Cities cases use models that are designed to simulate only the gross gain of only one category, while all the other cases use models that are designed to allow for simultaneous transitions among several categories. It is particularly challenging to write an algorithm for situations where more than one category competes to gain at a particular pixel.

8.3.2 *Modeling Challenges*

8.3.2.1 **To Separate Calibration from Validation**

Calibration is the procedure to set the parameters of a model, based on information at or before time 1. Validation is the procedure to assess how the predicted change compares to the reference change from time 1 to time 2. Proper validation of temporal prediction requires that calibration must be separate from validation though time. However, most of the cases used for calibration some information subsequent to time 1 in order to predict the change between time 1 and time 2. In 7 of the 13 cases, the model's calibration procedure used information directly from the reference map of time 2 concerning the quantity of each category. Other cases used influential variables, such as protected areas, that derive from contemporary time points subsequent to time 1. In these situations, it is impossible to determine whether the model's apparent accuracy indicates its predictive power through time. If a model uses information from both time 1 and time 2 for calibration, then the model's so-called prediction map of time 2 could be a match with the reference map of time 2 because the model parameters might be over fit to the data. The apparent accuracy would reflect a level of agreement higher than the level of agreement attributable to the model's predictive power into an unknown future.

There are some practical reasons why modelers use information subsequent to time 1 to predict the change between time 1 and time 2. Some reasons relate to the purpose of the model; other reasons relate to data availability.

The cases that applied LTM, CLUE-S and CLUE used information directly from the reference map of time 2 concerning the quantity of each category, because the priority for those applications was to predict the spatial allocation of land change. The user can specify the quantity of each category independently from the spatial allocation for these models, which can be an advantage in allowing them to be used with tabular data and other types of models that generate non-spatial information concerning only the quantity of each land category. For example, CLUE-S and CLUE can set the quantity of each category by using case-study-specific and scale-specific methods ranging from trend extrapolations to complex sectoral models of world trade.

Some models such as SAMBA require information that is available only for years after time 1. SAMBA is an agent-based modeling framework that uses information from interviews with farmers concerning their land practices. For the Cho Don case, these interviews were conducted subsequent to time 2. Furthermore, the purpose of the SAMBA model is to explore scenarios with local stakeholders, not to predict the precise allocation of land transitions. The SAMBA team has been developing other methods for process validation of various aspects of their model (Castella et al. 2005b; Castella and Verburg 2007).

There are costs associated with separating calibration from validation information, because strict separation prohibits the use of some variables that are known to influence land change but are available only for time points beyond the calibration time interval. The Worcester case accomplished separation between calibration information and validation information by restricting the use of independent variables. For example, maps of contemporary roads and protected areas are

available in digital form, but those maps contain some post-1971 information. The scientists for the Worcester application refrained from using these variables that are commonly associated with land change. Consequently, the Worcester case uses only slope and surficial geology as independent variables. Nevertheless, Pontius and Malanson (2005) show that there would not have been much increase in hits by using the map of protected areas, because such a map shows the places where change is prohibited, not the few places where change is likely to occur.

8.3.2.2   To Predict Small Amounts of Change

All 13 of the cases have less than 50% observed change, seven of the cases show less than 10% observed change, while the Holland(8), Santa Barbara, and Twin Cities have less than 4% observed change. Land change during a short time interval is usually a rare event, and rare events tend to be difficult to predict accurately. Figure 8.4 gives evidence that smaller amounts of change in the reference maps are associated with lower levels of predictive accuracy.

The challenge to detect and to predict change is made even more difficult by insisting upon rigorous separation of calibration data from validation data, especially in situations where data are scarce. For example, many models such as Environment Explorer are designed to examine change during a calibration interval from time 0 to time 1, and then to predict the change during a validation interval from time 1 to time 2. The Holland(15) case separates calibration information from validation information using this technique, where the calibration interval is only 7 years and the validation interval is only 4 years. In such situations, models may have difficulty in detecting a strong relationship between land change and the independent variables during the calibration interval, and the validation measurements may fail to find a strong relationship between the predicted land change and the observed land change during the validation interval. One solution would be for scientists to invest the necessary effort to digitize maps of historic land cover, so scientists can have a longer temporal extent and finer temporal resolution during which to calibrate and validate.

8.3.2.3   To Interpret the Influence of Quantity Error

Models that do not use the correct quantity of each category for time 2 must somehow predict the quantity for each category for time 2. Modelers need to be aware of how error in the prediction of quantity influences other parts of the validation process. Models typically fail to predict the correct allocation precisely; so models that predict more change are likely to produce more false alarms than models that predict less change, when assessed at fine spatial resolutions. For example, the Worcester case predicts more than the observed amount of change, which leads to false alarms. If the model were to predict less than the observed amount of change, then its output would have fewer false alarms and more correct rejections. In contrast, SLEUTH predicts less than half of the amount of observed change for the

Santa Barbara case, thus its error is close to that of a null model. It does not make sense to use criteria that reward systematic underestimates or overestimates of the quantity of each category. This is a weakness of using the percentage correct and the null model as benchmarks for predictive accuracy, and is a reason why Pontius et al. (2008) used the figure of merit as a criterion.

It is difficult to evaluate a model's prediction of spatial allocation when there is large error in quantity, especially when the model predicts less than the amount of observed change in the reference maps. We can assess the model's ability to predict spatial allocation somewhat when the model predicts the correct quantity, which is one reason modelers sometimes use the correct quantity at time 2 for simulation. Nevertheless, if we use only one potential realization of the model's output map, then the model's specification of spatial allocation is confounded with its single specification of quantity. The Total Operating Characteristic (TOC) is a quantitative procedure that can be used to measure a model's ability to specify the spatial allocation of land change in a manner that allows the modeler to consider various specifications of quantity (Pontius and Si 2014). Scientists can compute the TOC for cases where the model generates a map of relative priority for the gain of a particular category, which many models do in their intermediate steps. The TOC allows scientists to measure a model's ability to predict the few locations that change and a model's ability to predict the majority of locations that persist. The TOC is a recent advancement inspired by the Relative Operating Characteristic (Swets 1988; Pontius and Parmentier 2014).

8.3.3 Learning Challenges

8.3.3.1 To Use Appropriate Map Comparison Measurements

Scientists have invested a tremendous amount of effort to create elaborate algorithms to model landscape change. We are now at a point in our development as a scientific community to begin to answer the next type of question, specifically, "How well do these models perform and how do we communicate model performance to peers and others?" Therefore, we need useful measurements of map comparison and model performance. Pontius et al. (2008) derived a set of metrics to compare maps in a manner that we hope is both intellectually accessible and scientifically revealing, because analysis using rigorous and clear measurements is an effective way to learn. The initial invitation to participants asked them to submit their recommended criteria for map comparison. Few participants submitted any criteria, and those who did typically recommended the percentage of pixels in agreement between the reference map of time 2 and the prediction map of time 2.

This percentage correct criterion is one that many modelers consider initially. However, percentage correct can be extremely misleading, especially for cross-case comparisons. Percentage correct fails to consider the landscape dynamics, because percentage correct fails to include the reference map of time 1. For example, the Santa Barbara case has by far the largest percentage correct, 97%, simply because there is very little observed change on the landscape and the model predicts less

than the amount of observed change. On the other hand, the Cho Don case has the smallest percentage correct, 54%, primarily because the Cho Don case has more observed change than any other case. The Perinet case has the largest figure of merit, while its percentage correct of 81% ranks just below the median of the 13 cases. Producer's Accuracy, User's Accuracy, and Kappa are other indices of agreement that are extremely common in GIS and can be quite misleading in assessing the accuracy of land change models (Pontius and Millones 2011). The figure of merit has properties that are more desirable than metrics that are frequently used for pattern validation of land change models (Pontius et al. 2007, 2011). We recommend the figure of merit for situations when it is necessary to rank numerous model runs with a single measurement. However, a single measurement offers only one bit of information thus fails to convey various important aspects of a pattern validation. For example, the figure of merit fails to convey the size of the reference change relative to the size of the predicted change.

We recommend much more highly that modelers report the sizes of misses, hits, wrong hits and false alarms, which are the components of the figure of merit. That combination of four measures is helpful in a variety of respects. For example, the false alarms are fewer than the misses when the model predicts less change than the reference change; and the false alarms are more than the misses when the model predicts more change than the reference change. If there exist false alarms at some locations and misses at other locations, then there exists allocation error. It is helpful to distinguish allocation error from quantity error, because the two types of error can have different implications for practical interpretation depending on the model's purpose. For example, if the purpose of the model is to simulate total carbon dioxide emissions due to deforestation, then allocation error is less important than quantity error for spatial extents where forest biomass is homogeneous (Gutierrez-Velez and Pontius 2012).

We need to continue to invest effort to improve methods of map comparison. The Map Comparison Kit includes a variety of new tools (Visser and de Nijs 2006). Modules in the GIS software TerrSet allow scientists to compare maps where the pixels have simultaneous partial membership to several categories, which is essential for multiple resolution comparison (Pontius and Connors 2009). The free software R contains packages that land change scientist will find helpful. The TOC package computes the Total Operating Characteristic (Pontius and Si 2014). The diffeR package gives components of difference at multiple spatial resolutions for two maps that show a single variable, such as maps from times 1 and 2 (Pontius and Santacruz 2014). Moulds et al. (2015) created in R the lulcc package, which performs a variety of operations, including the multiple resolution calculation of misses, hits, wrong hits, false alarms and correct rejections as derived by Pontius et al. (2011).

8.3.3.2 To Learn About Land Change Processes

During the panel discussions, participants agreed that a main purpose of modeling land use and cover change (LUCC) is to increase understanding of processes of LUCC, and that scientists should design a research agenda in order to maximize

learning concerning such processes, not merely to increase predictive accuracy. Therefore, scientists should strive to glean from a validation exercise useful lessons about the processes of land change and about the next steps in the research agenda.

Some attendees at the panel discussions expressed concern that this chapter's validation exercises focus too much on prediction to the exclusion of increasing our understanding of the underlying processes of LUCC. Many scientists profess to seek explanation, not necessarily prediction. Some scientists think that a model can predict accurately for the wrong reasons; in addition these scientists think a model can capture the general LUCC processes, but not necessarily predict accurately due to inherent unpredictability of the processes. These participants reminded the audience that pattern validation examines the output maps from the simulation models but does not examine whether the structure of the algorithm matches theory concerning the processes of change. Process validation is required to validate the structure of the algorithm for process based models, especially when path dependence plays a role (Brown et al. 2005).

Other scientists see pattern validation as a means to distinguish better explanations from poorer explanations concerning the LUCC processes. For these other scientists, pattern validation allows a modeler to gain insight concerning the degree to which the simulated change is similar to the observed change. Furthermore, scientists must test the degree to which the past is useful to predict the future because this allows scientists to measure the scales at which LUCC processes are stable over time. A model's failure to predict accurately may indicate that the process of land change is non-stationary in time and/or space, in which case pattern validation can reveal information that is helpful to learn about LUCC processes (Chen and Pontius 2010; Pontius and Neeti 2010). Thus there is need for new methods, such as Intensity Analysis, that test for stationarity at various levels, even before any predictive model is run (Aldwaik and Pontius 2013; Runfola and Pontius 2013). If scientists interpret the validation procedure in an intelligent manner, then they can perhaps learn more from inaccurate predictions than from accurate ones. Consequently, inaccurate predictions do not mean that the model is a failure, because validation can lead to learning regardless of the revealed level of accuracy.

This difference in views might explain the variation in the LUCC modeling community concerning how best to proceed. One group thinks that models are too simple so that future work should consider more variables and develop more complex algorithms so the models can generate a multitude of possible outcomes. A second group insists that such an approach would only exacerbate an existing problem that models are already too complicated to allow for clear communication, even among experts. From this second perspective, contemporary models lack aspects of scientific rigor that would not be corrected by making the models more complex. For example, many models fail to separate calibration information from validation information, fail to apply useful methods of map comparison, and fail to measure how scale influences the analysis. For this second group of scientists, it would be folly to make more complicated algorithms and to include more variables before we tackle basic issues, because we will not be able to measure whether more complex models actually facilitate learning about LUCC processes until we develop and use

helpful measures of model performance. This apparent tension could be resolved if the scientists who develop more complex models collaborate with the scientists who develop clearer methods of model assessment.

8.3.3.3 To Collaborate Openly

Participants at the panel sessions found the discussions particularly helpful because the sessions facilitated open and frank cross-laboratory communication. Many conference participants expressed gratitude to the co-authors who submitted their maps in a spirit of openness for the rest of the community to analyze in ways that were not specified a priori. The design of the exercise encouraged participation and open collaboration because it was clear to the participants that the analysis was not attempting to answer the question "Which model is best?"

Some participants in the conference discussions reported that they have felt professional pressure to claim that their models performed well in order for their manuscripts to be accepted for publication in peer-reviewed journals. We hope that this chapter opens the door for honest and helpful reporting about modeling results. In particular, we hope that editors and reviewers will learn as much from this study as the conference participants did, so that future literature includes useful information about model assessment. The criterion for acceptance of manuscripts should be rigor of method and clarity of presentation, not results concerning predictive accuracy, and certainly not vacuous claims of success.

There is clearly a desire to continue this productive collaboration because it greatly increases learning. One particularly constructive suggestion is to build a LUCC data digital library so that scientists would have access to each others' data, models, and modeling results. The data would be peer-reviewed and have metadata sufficient so that anyone could perform cross-model comparison with any of the entries in the library. In order for this to be successful, scientists need sufficient motivation to participate, which requires funding and professional recognition for participation.

8.4 Conclusions

The collective experience of the co-authors supports the statement that all models are wrong but some are useful (Box 1979). All 13 of the models are wrong in the respect that the outputs have errors. Errors in pattern validation mean that the patterns extrapolated from the calibration time interval were not stationary with the patterns observed during the validation time interval. These errors are a reflection of the landscape as much as they are a reflection of the model. If the scientists interpret the results in a useful manner, then scientists can learn; and if scientists learn from a model, then the model was successful at advancing science. It is essential to use measurements that can be interpreted with respect to a model's intended purposes in

order to facilitate learning. Clarity and rigor are necessary to establish procedures and measurements for informative judgments concerning model performance. This chapter illuminates common pitfalls and offers guidance for ways to overcome the pitfalls. Specifically, we recommend modelers report the sizes of misses, hits, wrong hits, and false alarms. Those four measurements are based on the mathematical ideas concerning the intersection of sets, which are regularly taught to elementary school students. If scientists meet the challenges specified in this chapter, then we are likely to learn efficiently, because meeting these challenges can help scientists prioritize a research agenda for land change science. To facilitate open collaboration, we have made the raster maps used in this cross-case comparison available for free at www.clarku.edu/~rpontius.

References

Aldwaik SZ, Pontius RG Jr (2013) Map errors that could account for deviations from a uniform intensity of land change. Int J Geogr Inf Sci. doi:10.1080/13658816.2013.787618

Aldwaik SZ, Onsted JA, Pontius RG Jr (2014) Behavior-based aggregation of land categories for temporal change analysis. Int J Appl Earth Obs Geoinf 35:229–238

Boissau S, Castella J-C (2003) Constructing a common representation of local institutions and land use systems through simulation-gaming and multi-agent modeling in rural areas of Northern Vietnam: the SAMBA-Week methodology. Simul Gaming 34(3):342–347

Box GEP (1979) Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN (eds) Robustness in statistics. Academic, New York, pp 201–236

Brown DG, Page S, Riolo R, Zellner M, Rand W (2005) Path dependence and the validation of agent-based spatial models of land use. Int J Geogr Inf Sci 19(1):153–174

Brown, D.G., Band, L.E., Green, K.O., Irwin, E.G., Jain, A., Lambin, E.F., Pontius Jr, R.G., Seto, K.C., Turner II, B.L., Verburg, P.H. (2013). Advancing land change modeling: opportunities and research requirements. The National Academies Press: Washington, DC. 145. http://www.nap.edu/catalog.php?record_id=18385

Castella J-C, Verburg PH (2007) Combination of process-oriented and pattern-oriented models of land-use change in a mountain area of Vietnam. Ecol Model 10(1):410–420

Castella J-C, Boissau S, Trung TN, Quang DD (2005a) Agrarian transition and lowland-upland interactions in mountain areas in northern Vietnam: application of a multi-agent simulation model. Agric Syst 86(3):312–332

Castella J-C, Trung TN, Boissau S (2005b) Participatory simulation of land use changes in the Northern Mountains of Vietnam: the combined use of an agent-based model, a role-playing game, and a geographic information system. Ecol Soc 10(1):27

Chen H, Pontius RG Jr (2010) Diagnostic tools to evaluate a spatial land change projection along a gradient of an explanatory variable. Landsc Ecol 25:1319–1331

de Koning GHJ, Verburg PH, Veldkamp TA, Fresco LO (1999) Multi-scale modelling of land use change dynamics in Ecuador. Agric Syst 61:77–93

de Nijs TCM, de Niet R, Crommentuijn L (2004) Constructing land-use maps of the Netherlands in 2030. J Environ Manag 72(1–2):35–42

Dietzel CK, Clarke KC (2004) Spatial differences in multi-resolution urban automata modeling. Trans GIS 8:479–492

Duan Z, Verburg PH, Fengrong Z, Zhengrong Y (2004) Construction of a land-use change simulation model and its application in Haidian District, Beijing. Acta Geograph Sin 59(6):1037–1046. (in Chinese)

Enaruvbe G, Pontius RG Jr (2015) Influence of classification errors on intensity analysis of land changes in southern Nigeria. Int J Remote Sens 31(1):244–261

Engelen G, White R, de Nijs T (2003) The Environment Explorer: spatial support system for integrated assessment of socio-economic and environmental policies in the Netherlands. Integr Assess 4(2):97–105

Fotsing E, Verburg PH, De Groot WT, Cheylan J-P, Tchuenté M (2013) Un modèle intégré pour explorer les trajectoires d'utilisation de l'espace. ARIMA J 16:1–28

Goldstein NC (2004) Brains vs. Brawn – comparative strategies for the calibration of a cellular automata-based urban growth model. In: Atkinson P, Foody G, Darby S, Wu F (eds) GeoDynamics. CRC Press, Boca Raton, pp 249–272

Gutierrez-Velez V, Pontius RG Jr (2012) Influence of carbon mapping and land change modelling on the prediction of carbon emissions from deforestation. Environ Conserv 39(4):325–336

Hilferink M, Rietveld P (1999) Land use scanner: an integrated GIS based model for long term projections of land use in urban and rural areas. J Geogr Syst 1(2):155–177

Hoymann J (2010) Spatial allocation of future residential land use in the Elbe River Basin. Environ Plan B: Plan Des 37(5):911–928

Kok K, Veldkamp TA (2001) Evaluating impact of spatial scales on land use pattern analysis in Central America. Agric Ecosyst Environ 85(1–3):205–221

Kok K, Farrow A, Veldkamp TA, Verburg PH (2001) A method and application of multi-scale validation in spatial land use models. Agric Ecosyst Environ 85(1–3):223–238

Koomen, E, Borsboom-van Beurden, J. (2011). Land-use modelling in planning practice. GeoJ Libr 101, Dordrecht: Springer

Kuhlman T, Diogo V, Koomen E (2013) Exploring the potential of reed as a bioenergy crop in the Netherlands. Biomass Bioenergy 55:41–52

Laney RM (2002) Disaggregating induced intensification for land-change analysis: a case study from Madagascar. Ann Assoc Am Geogr 92(4):702–726

Loonen W, Koomen E (2009) Calibration and validation of the land use scanner allocation algorithms, PBL publication number 550026002. Netherlands Environmental Assessment Agency (PBL), Bilthoven

McConnell W, Sweeney SP, Mulley B (2004) Physical and social access to land: spatio-temporal patterns of agricultural expansion in Madagascar. Agric Ecosyst Environ 101(2–3):171–184

Moulds S, Buytaert W, Mijic A (2015) An open and extensible framework for spatially explicit land use change modelling: the lulcc R package. Geosci Model Dev 8:3215–3229

Paegelow M, Camacho Olmedo MT, Houet T, Mas J-F, Pontius RG Jr (2013) Land change modeling: moving beyond projections. Int J Geogr Inf Sci 27(9):1691–1695

Pijanowski BC, Gage SH, Long DT (2000) A land transformation model: integrating policy, socioeconomics and environmental drivers using a geographic information system. In: Harris L, Sanderson J (eds) Landscape ecology: a top down approach. CRC Press, Boca Raton, pp 183–198

Pijanowski BC, Brown DG, Manik G, Shellito B (2002) Using neural nets and GIS to forecast land use changes: a land transformation model. Comput Environ Urban Syst 26(6):553–575

Pijanowski BC, Pithadia S, Sellito BA, Alexandridis K (2005) Calibrating a neural network-based urban change model for two metropolitan areas of the Upper Midwest of the United States. Int J Geogr Inf Sci 19(2):197–215

Pontius RG Jr, Connors J (2009) Range of categorical associations for comparison of maps with mixed pixels. Photogramm Eng Remote Sens 75(8):963–969

Pontius RG Jr, Lippitt CD (2006) Can error explain map differences over time? Cartogr Geogr Inf Sci 33(2):159–171

Pontius RG Jr, Malanson J (2005) Comparison of the structure and accuracy of two land change models. Int J Geogr Inf Sci 19(2):243–265

Pontius RG Jr, 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 Jr, Neeti N (2010) Uncertainty in the difference between maps of future land change scenarios. Sustain Sci 5:39–50

Pontius RG Jr, Parmentier B (2014) Recommendations for using the Relative Operating Characteristic (ROC). Landsc Ecol 29(3):367–382

Pontius RG Jr, Petrova S (2010) Assessing a predictive model of land change using uncertain data. Environ Model Softw 25(3):299–309

Pontius RG Jr, Santacruz A (2014) Quantity, exchange and shift components of differences in a square contingency table. Int J Remote Sens 35(21):7543–7554

Pontius RG Jr, 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 Jr, Cornell J, Hall C (2001) Modeling the spatial pattern of land-use change with GEOMOD2: application and validation for Costa Rica. Agric Ecosyst Environ 85(1–3):191–203

Pontius RG Jr, Huffaker D, Denman K (2004a) Useful techniques of validation for spatially explicit land-change models. Ecol Model 179(4):445–461

Pontius RG Jr, Shusas E, McEachern M (2004b) Detecting important categorical land changes while accounting for persistence. Agric Ecosyst Environ 101(2–3):251–268

Pontius RG Jr, Walker R, Yao-Kumah R, Arima E, Aldrich S, Caldas M, Vergara D (2007) Accuracy assessment for a simulation model of Amazonian deforestation. Ann Assoc Am Geogr 97(4):677–695

Pontius RG Jr, 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, Pijanowski 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):11–47

Pontius RG Jr, Peethambaram S, Castella J-C (2011) Comparison of three maps at multiple resolutions: a case study of land change simulation in Cho Don District, Vietnam. Ann Assoc Am Geogr 101(1):45–62

Runfola D, Pontius RG Jr (2013) Measuring the temporal instability of land change using the flow matrix. Int J Geogr Inf Sci. doi:10.1080/13658816.2013.792344

Silva EA, Clarke KC (2002) Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. Comput Environ Urban Syst 26:525–552

Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293

Tan M, Li X, Xie H, Lu C (2005) Urban land expansion and arable land loss in China − a case study of Beijing-Tianjin-Hebei region. Land Use Policy 22(3):187–196

Veldkamp AT, Fresco L (1996) CLUE-CR: an integrated multi-scale model to simulate land use change scenarios in Costa Rica. Ecol Model 91:231–248

Verburg PH, Veldkamp TA (2004) Projecting land use transitions at forest fringes in the Philippines at two spatial scales. Landsc Ecol 19:77–98

Verburg PH, de Koning GHJ, Kok K, Veldkamp A, Bouma J (1999) A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. Ecol Model 116(1):45–61

Verburg PH, Soepboer S, Veldkamp TA, Limpiada R, Espaldon V, Sharifah Mastura SA (2002) Modeling the spatial dynamics of regional land use: the CLUE-S model. Environ Manag 30(3):391–405

Verburg PH, de Nijs TCM, van Eck JR, Visser H, de Jong K (2004) A method to analyse neighbourhood characteristics of land use patterns. Comput Environ Urban Syst 28(6):667–690

Visser H, de Nijs T (2006) The map comparison kit. Environ Model Softw 21(3):346–358