Environmental Modelling & Software 53 (2014) 121–136

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Identifying a land use change cellular automaton by Bayesian data assimilation

Judith A. Verstegen [a,*], Derek Karssenberg [b], Floor van der Hilst [a], André P.C. Faaij [a]

[a] Copernicus Institute for Sustainable Development and Innovation, Faculty of Geosciences, Utrecht University, Heidelberglaan 2, 3584 CS Utrecht, The Netherlands

[b] Department of Physical Geography, Faculty of Geosciences, Utrecht University, Heidelberglaan 2, PO Box 80115, 3508 TC Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 25 January 2013
Received in revised form
15 November 2013
Accepted 28 November 2013
Available online 20 December 2013

Keywords:
Data assimilation
Cellular automata
Calibration
Model structure
Land use change
Particle filter

ABSTRACT

We present a Bayesian method that simultaneously identifies the model structure and calibrates the parameters of a cellular automaton (CA). The method entails sequential assimilation of observations, using a particle filter. It employs prior knowledge of experts to define which processes might be important in the system, and uses empirical information from observations to identify which ones really are and how these processes should be parameterized. In a case study for the São Paulo state in Brazil, we identify a land use change CA simulating sugarcane cropland expansion from 2003 to 2016. Eight annual observation maps of sugar cane cultivation are used, split over space and time for calibration and validation. It is shown that the identified CA can properly reproduce the observations, and has a minimum reduction factor of 3 in root mean square error compared to a Monte Carlo simulation without particle filter. In the part of the study area where no observational data are assimilated (validation area), there is little reduction in model performance compared to the part with observational data. So, incomplete datasets, regional land survey data, or clouded remote sensing images can still provide useful information for this particle filter method, which is an advantage because good quality land use maps are rare. Another advantage is that in our approach the output uncertainty encompasses errors from expert knowledge, model structure, parameters and observation (calibration) data. This can, in our opinion, be very useful for example to determine up to what future period the results are a secure basis for decisions and policy making.

1. Introduction

A Cellular Automaton (CA) represents spatio-temporal change as local interactions of different entities and processes in a raster environment (Santé et al., 2010). The fact that a CA consists of relatively simple rules that can lead to complex patterns, makes it suitable to study complex system behaviour, which is currently considered important in environmental systems research (Page, 2011; Manson, 2007; Johnson, 2010; Grimm and Railsback, 2012). Therefore, cellular automata are applied in many environmental modelling domains, like fire propagation (Berjak and Hearne, 2002), vegetation spreading (Kéfi et al., 2007), and urban or land use change modelling (Verburg et al., 2004; Batty, 2005; Lauf et al., 2012). In CA development, one can distinguish between model structure identification, i.e. finding the set of processes to be represented in the model, conceptualized into the set of transition rules, and model calibration, i.e. finding the correct parameterization of these processes. In urban and land use change modelling, finding the set of transition rules is problematic (Santé et al., 2010; Straatman et al., 2004), which possibly poses limitations on the reliability and therefore the usability of these models.

Transition rule derivation can be done in a number of ways. 1) From fundamental, e.g., physical or chemical, laws (e.g., Collin et al., 2011). This is difficult in land use change modelling, as most fundamental laws in this field do not provide a quantitative process description. Yet, some have successfully applied physical laws to simulate land use expansion, mainly aimed at cities (Batty, 2012; Bettencourt, 2013). 2) By experts, who have experience-based knowledge of the study area. This is widely done in land use change modelling (e.g., van der Hilst et al., 2012; Yu et al., 2011), but it is somewhat subjective. 3) From empirical data. It is recognized that this is challenging in land use change modelling (Straatman et al., 2004; Hansen, 2012), but it is still important to continue exploring this option, because there is a need to find a more evidence-based approach to set up a land use change model.

* Corresponding author.
E-mail address: J.A.Verstegen@uu.nl (J.A. Verstegen).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.11.009

One can combine the benefits of expert knowledge and empirical data by using a method for transition rule derivation in which the prior knowledge (definition of potential model structures) is defined by experts, and the posterior knowledge (identification of the best structure) is attained by empirical data. Our objective is to devise such a method, which we believe should fulfil two requirements. The first requirement is that the method should be able to quantify uncertainty (Rasmussen and Hamilton, 2012; Aerts et al., 2003), i.e. it should not only be able to select the best model structure from all potential model structures defined by the prior knowledge, but it should give the likelihood of each individual structure being correct. In this way, a stochastic CA is obtained, which combines all potential model structures and parameters in an optimal way. The most important advantage of this is that confidence intervals of the modelled land use projections can be defined, such that policy makers can decide up to what point in time the projections are reliable enough to be a foundation for their policies. The second requirement is that, herein, one should not only take into account uncertainty in the prior information, but also in the empirical data, the observations of land use, used to update the priors (Fang et al., 2006). Ignoring uncertainty in the empirical data may lead to an underestimation of model output uncertainty.

The combined requirements of prior knowledge, observation uncertainty, and posterior knowledge with output uncertainty lead towards Bayesian methods, which start out with prior knowledge, and then assemble model uncertainty and observation uncertainty to end up with posterior knowledge including uncertainty information. Therefore, we show a method for model structure identification and calibration using the particle filter, a sequential Bayesian estimation, or data assimilation, technique (van Leeuwen, 2009). Data assimilation techniques update the prior knowledge during model runtime at time steps when observations are available. We will use this property to sequentially update both the model rules and their parameters. Data assimilation techniques are increasingly being used to calibrate spatio-temporal models in a wide range of different fields in the environmental sciences, such as oceanography (van Leeuwen, 2003), hydrology (Salamon and Feyen, 2009), and atmospheric transport (Hiemstra et al., 2012), but have, to our knowledge, not yet been applied for model structure identification. Recently, their potential has been recognized in the land use change field (van der Kwast et al., 2011; Zhang et al., 2011).

The approach that is most often used in land use change modelling to define the model structure is regression on a land use map (Verburg et al., 1999, 2002; Aguiar et al., 2007; Diogo et al., in preparation). This method mostly results in only one deterministic model structure, without uncertainty in either the observations used to construct the regression model or in the model itself, and therefore does not meet our requirements. In the last decade, model rule identification methods originating from artificial intelligence have become popular, like neural networks (Dai et al., 2005; Li and Yeh, 2002), and swarm intelligence algorithms (Feng et al., 2011; Liu et al., 2008). These, however, do not take into account observation uncertainty, the second requirement. Moreover, they result in black-box models (Li and Yeh, 2002), i.e. they do not provide explicit posterior knowledge. Bayesian land use model structure identification has been performed before by Kocabas and Dragićević (2007). They apply a Bayesian network and an influence diagram. However, they do not include observation uncertainty.

In this study, we evaluate the performance of the particle filter method for model structure identification and calibration of a land use change CA. Furthermore, we assess the effect of the amount of observational data assimilated, because time series of good quality land use maps are often absent (Straatman et al., 2004). We also consider the effect of a pre-set (expert-based) model structure, to represent the situation of a model structure identification determined beforehand, which is now common practice in land use change modelling. In all approaches we provide confidence intervals with the land use projections, useful as a decision criterion for policy makers.

The assessments are carried out on a case study of the expansion of sugar cane fields in the São Paulo state in Brazil, using an adapted form of the PCRaster Land Use Change model (PLUC) (Verstegen et al., 2012). As the sugar cane is partly used to produce ethanol, this case study is relevant in view of the current debate on the sustainability of bioenergy from dedicated crops when land use change is taken into account (Lapola et al., 2010; Hellmann and Verburg, 2011). São Paulo is especially interesting because it has a long history in ethanol production (Walter et al., 2011) and very good observational data availability (Rudorff et al., 2010).

The next section provides a definition of the problem of transition rule identification in a CA, a brief explanation of data assimilation, a description of the case study, an outline of the prior information about the land use change model structure and parameters, details of the performance measures used, a description of the observational data, and a scenario sketch. This is followed by a combined results and discussion section, and a conclusion section.

2. Methods

2.1. Model structure and parameter identification in a land use change cellular automaton

A cellular automaton (CA) consists of a set of transition rules representing the processes that lead to change in the system state over time and rules to combine these transition rules (Fig. 1). In the case of a land use change CA a transition rule is a function calculating the suitability of each location (cell) for a particular land use type, with respect to a spatial attribute that influences the allocation of that land use type, for instance the slope or the distance to roads. So, a land use change CA contains for each land use type a set of transition rules. The transition rules contain parameters defining the characteristics of the process represented by the transition rule, for example an exponent in an exponential relationship between the suitability value and slope. The transition rules need to be selected and combined such that they represent the key processes that steer the spatial allocation of land use change. This can be accomplished by selecting from a set of candidate transition rules. This could be done either in a Boolean fashion, by switching transition rules on or off, or in a continuous fashion, by weighting each transition rule. We refer to this selection of transition rules as model structure identification.

In modelling, it is essential to find the model structure and parameter values that result in an optimal model representation of the studied land use system. Identification of the model structure and parameter values can be accomplished through comparison of the modelled system, with certain transition rules and

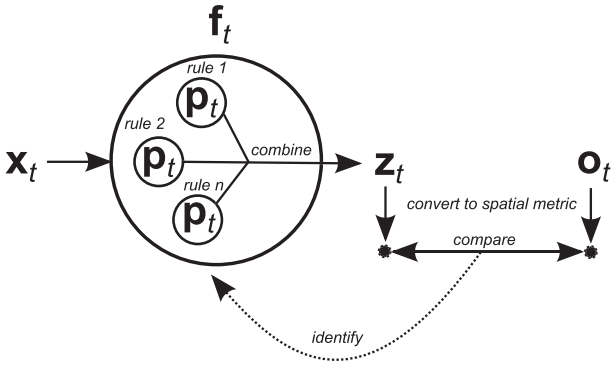

Fig. 1. Conceptual model of a general CA: represents the processes of change in the system state over time, i.e. the set of transition rules and the way to combine them, $\mathbf{x}_t$ represents all inputs, usually spatial attributes, and contains the parameters. Model calibration refers to identifying $\mathbf{p}_t$, model structure identification refers to selecting the transition rules. Identification of the parameters and model structure is based on a comparison between the land use map $\mathbf{z}_t$, or a derived spatial metric, with the observed land use map $\mathbf{o}_t$, or a derived spatial metric. The parameter values and model structure with the smallest difference between $\mathbf{z}_t$ and $\mathbf{o}_t$ are considered optimal.

parameter values, and observations of the real system (Fig. 1, right side), subsequently selecting the parameter values and model structure that minimize the difference between modelled and observed land use. Parameter identification, or calibration, has become common practice in land use change CA modelling, although the applied method differs per study (Santé et al., 2010). But methods to identify the transition rules, or model structure, are generally lacking (Straatman et al., 2004). Here, we propose a technique to simultaneously identify the parameters and the model structure using observational data.

To summarize, a general CA, with the system state variable(s) $\mathbf{z}_t$ and initial state(s) $\mathbf{z}_0$, can be defined as:

$$\mathbf{z}_t = \mathbf{f}_t(\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{p}_t), \quad \text{for each } t = 1, 2, \ldots, \mathrm{T} \tag{1}$$

In Equation (1), $\mathbf{f}_t$ is the set of transition rules at time step t, representing the processes that lead to change in the system state over time. The vector $\mathbf{x}_t$ represents all inputs, usually spatial attributes, and boundary conditions and $\mathbf{p}_t$ contains the parameters. In a stochastic model, the uncertain parts of the system are described stochastically. So, we have $p(\mathbf{f}_t)$, a probability distribution of possible transition rules, $p(\mathbf{z}_{t-1})$ the probability distribution of the previous system states, $p(\mathbf{x}_t)$, the probability distribution of inputs and boundary conditions, and $p(\mathbf{p}_t)$, the probability distribution of the parameters. In the case that no observational data are used, these distributions together determine the shape of the resulting probability distribution of the state variable, referred to as $p(\mathbf{z}_t)$. Yet, our aim is to use observations to simultaneously identify $p(\mathbf{f}_t)$ and $p(\mathbf{p}_t)$ in such a way that the model output matches the observations as closely as possible.

2.2. General particle filter framework

If we want to use the information comprised in system observations $\mathbf{o}_t$ to select the transition rules (model structure identification) and parameterizations (calibration), that perform well and to incorporate this knowledge into the model, $p(\mathbf{z}_t)$ should be updated (the 'identify' step in Fig. 1). Bayes' rule updates a probability distribution of a variable, when evidence, i.e. an observation, of this variable arrives. So, for the time steps at which observational data are available the following equation is evaluated.

$$p(\mathbf{z}_t|\mathbf{o}_t) = \frac{p(\mathbf{o}_t|\mathbf{z}_t) \cdot p(\mathbf{z}_t)}{p(\mathbf{o}_t)}, \quad \text{for each } t \tag{2}$$

In Equation (2), $p(\mathbf{o}_t)$ is the probability distribution of the observations, i.e. the measurement data and their uncertainty. Thereby, model structure identification using Bayes' rule fulfils our second requirement of taking into account observation uncertainty. $p(\mathbf{o}_t|\mathbf{z}_t)$ is the joint probability density of the observations at t given the model state, which can be seen as the likelihood that the observations occur given the model. The posterior probability $p(\mathbf{z}_t|\mathbf{o}_t)$ is the probability distribution of the state variable $p(\mathbf{z}_t)$ adjusted to the observations. Hence, Bayes' rule quantifies output (posterior) uncertainty given observation and input (prior) uncertainty, thereby satisfying our first requirement.

Numerically, Equation (1) is often solved using Monte Carlo analysis, which represents probability distributions by a number of realizations, N, of the model. Several sequential data assimilation techniques are available to solve Equation (2). Most data assimilation techniques are based on filtering theory (Jazwinksi, 1970): they filter the Monte Carlo realizations sequentially over time. The most well-known filter technique is probably the Ensemble Kalman filter, first introduced by Evensen (1994). This filter is, however, not guaranteed to work with non-Gaussian distributions and non-linear systems and is thus not suitable for identifying transition rules in complex systems (Pasetto et al., 2012). More importantly, the standard version of the Ensemble Kalman filter allows updates only for the variables for which observations are available, which makes updating model structure and parameters in a land use change CA impossible, as these are not observable in the real world. Versions of the Ensemble Kalman filter that do allow this require very strict premises that do not hold for cellular automata. Therefore, we have selected another data assimilation technique that allows updating all, also non-observed, model variables and can handle non-Gaussianity and non-linearity: the sequential importance resampling (SIR) particle filter (van Leeuwen, 2009), hereafter simply referred to as the particle filter. Here, we only provide a short description of the particle filter. For a more extensive introduction into the particle filter, see e.g., Arampulam et al. (2002), Bengtsson et al. (2008), and van Leeuwen (2009). At each time step for which observational data are available the particle filter uses Bayes' rule (Equation (2)) to assess the probability that a certain Monte Carlo realization, here called particle, and the observed data can be considered equal (Hartig et al., 2011). Herein, the following steps are taken (Fig. 2):

1. N realizations are drawn from the initial probability distributions of model structures $\mathbf{f}_t$, inputs $\mathbf{x}_t$, and parameters $\mathbf{p}_t$ (Equation (1)), resulting in a total number of N particles.
2. For all N particles the land use change model (explained in Section 2.3.2) is run up to the next filter moment, i.e. the next moment for which observational data are available.

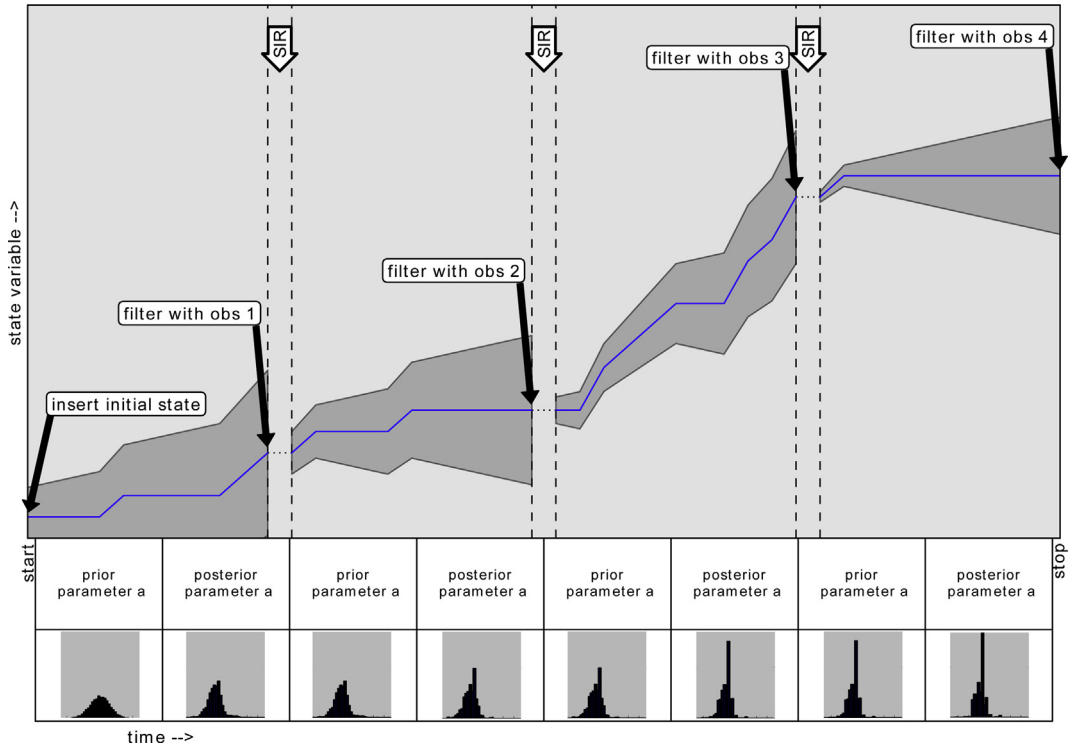

Fig. 2. Functioning of the particle filter. 'Obs 1' means observations at filter moment 1, the blue line indicates the median system state, grey areas represent the confidence interval. Histograms underneath the plots illustrate the effect of the filter moments on a parameter, referred to as parameter a. The prior distribution of parameter a at filter moment t is always equal to the posterior of parameter a at filter moment $t-1$. The effect on the transition rules is the same (not shown).

3. The posterior probability that the modelled state at that moment is correct given the observations with their uncertainty (explained in Section 2.3.4), is calculated for each of the particles. A posterior probability of one indicates a perfect match and a posterior probability of zero a complete mismatch.
4. Now, the sequential importance resampling (SIR) is performed: N particles are drawn to be progressed to the next observation moment with probabilities that are proportional to the posterior probabilities calculated in step 3. This procedure causes particles with a high posterior probability to be copied often (drawn several times) and particles with a low posterior probability to be removed (never drawn).
5. Steps 2 to 4 are repeated until all filter moments are completed and the model has reached the final time step. This means that at each filter moment the initial distributions obtained in step 1 are narrowed, i.e. the number of unique particles diminishes over time (see e.g., the histograms of parameter a in Fig. 2).

Note that, whenever a particle is copied, all model components within it are copied. Within $p(\mathbf{z}_t|\mathbf{o}_t)$ the transition rules $p(\boldsymbol{f}_t)$ and parameters $p(\mathbf{p}_t)$ are thereby updated as well. So, after assimilation of all observations, i.e. after the last filter moment, the best model structure is identified and the CA is calibrated, i.e. the posterior probability distributions of $\mathbf{f}_t$ and $\mathbf{p}_t$ are obtained.

For conducting these steps, the PCRaster Python framework is used, which is freely available via http://pcraster.geo.uu.nl (Karssenberg et al., 2010). Steps 1 and 2 involve running the land use change model in Monte Carlo mode, as explained in Verstegen et al. (2012). Step 3 is achieved by solving Bayes' theorem for each particle:

$$p\left(\mathbf{z}_t^i|\mathbf{o}_t\right) = \frac{p(\mathbf{o}_t|\mathbf{z}_t^i)\cdot p(\mathbf{z}_t^i)}{\sum_{j=1}^{N} p\left(\mathbf{o}_t|\mathbf{z}_t^i\right)\cdot p\left(\mathbf{z}_t^i\right)}, \text{ for each } i = 1,2,\ldots,\mathrm{N} \tag{3}$$

In Equation (3), $p(\mathbf{z}_t^i)$ is the prior probability of model realization i, which is always equal to $1/N$ because the same number of particles is drawn at each filter moment (step 4). If the observations are not of the state variable, but of a derived spatial metric, like relative proportions of land use in a subarea as it is often found in census data, the model state $\mathbf{z}_t^i$ has to be converted to that measure before filtering.

In Equation (3), $p(\mathbf{z}_t^i|\mathbf{o}_t)$ is the posterior probability of particle i and $p(\mathbf{o}_t|\mathbf{z}_t^i)$ is the probability of the observations given particle i. Under the assumption that the observation error has a Gaussian distribution, the latter can be calculated as (van Leeuwen, 2009):

$$p\left(\mathbf{o}_t|\mathbf{z}_t^i\right) = e^{-\frac{1}{2}\left[\mathbf{o}_t-\mathbf{z}_t^i\right]^T \mathbf{R}_t^{-1}\left[\mathbf{o}_t-\mathbf{z}_t^i\right]}, \text{ for each } t \tag{4}$$

In Equation (4), $\mathbf{R}_t$ is the covariance matrix of the observation error and T indicates matrix transposition. Going through steps 1–5, the procedure 'filters' the ensemble of particles because many particles do not match the observations, receive low weights, and are thus not drawn and not progressed to the next observation moment. So, although the number of particles remains the same, due to the resampling in step 4, the variation in the particles in terms of their uniqueness in the transition rules and parameters diminishes. This means that the initial probability distributions of these model components are narrowed. Hence, the particle filter has identified which transition rules are most likely to be valid (model structure), and in what ranges the parameters are most likely to fall. The model has thereby been calibrated.

2.3. Identifying transition rules of a land use change CA

2.3.1. Case study

A case study is defined to test the usability of the particle filter for model structure identification and calibration of a land use change CA. An important current debate in the land use change domain is whether bioenergy from dedicated crops is still sustainable when land use change (direct and indirect) is taken into account, in view of, e.g., carbon emissions (Lapola et al., 2010; Fargione et al., 2008; Searchinger et al., 2008), rising food prices (von Braun, 2008), and biodiversity (Hellmann and Verburg, 2010). For all these aspects it is important to know where bioenergy crops have expanded in the past and are likely to expand in the future. Such projections can be made with a land use change CA.

A key player in the bioenergy market is Brazil, mainly with the production of ethanol from sugar cane. Within Brazil, the state of São Paulo (Fig. 3) has the longest history as well as the largest share in sugar cane production (about 60% of the national production in recent years). In addition it still experiences a significant production growth (Walter et al., 2011; Sparovek et al., 2009), especially since the introduction of the flex-fuel car in 2003 (Macedo and Seabra, 2008). The actuality of the debate, together with availability of an annual spatial dataset of sugarcane cropland distribution from the Canasat project of the National Institute for Space Research in Brazil (INPE) (Rudorff et al., 2010) as observational data, makes sugar cane cropland expansion in the São Paulo state a suitable case for testing the merits of the particle filter for identifying a CA.

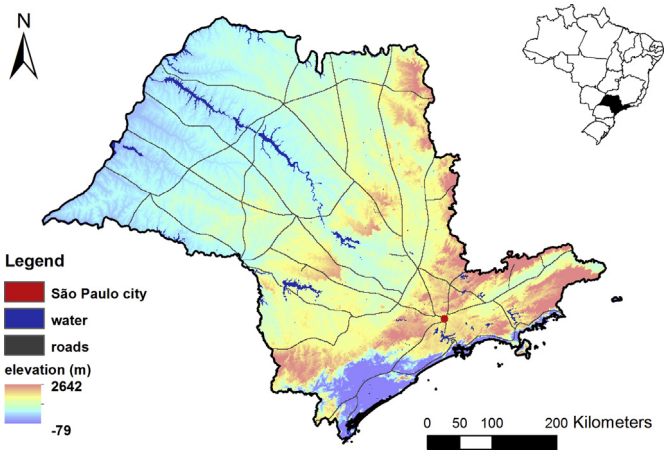

Fig. 3. Study area.

2.3.2. Transition rules and parameters: prior information

For simulating sugar cane cropland expansion an adapted form of the PCRaster Land Use Change model (PLUC) (Verstegen et al., 2012) is used. The land use change transition function, $\mathbf{f}_t$ in Equation (1), is regulated by the spatial allocation mechanism, which relies on the land demand and the suitability map (Fig. 4).

Different allocation mechanisms are possible, involving various degrees of competition between land use types. For instance, a model can fulfil the allocation of all land use types one by one (first the complete land claim of one land use type and then that of the next land use type), or assign to each cell the land use type with the highest suitability in that cell, or use a stochastic variant of the latter method. We use, however, a fixed, deterministic allocation mechanism, meaning that the cell with highest suitability is allocated first. This can be justified by the fact that there is little or no difference between different allocation mechanisms when considering only a single land use type.

The amount of land that is allocated or removed is steered by the demand (d_t) for products associated with the land use types, in this case sugar cane. In the current study, demand is expressed in hectares of cultivated land. All maps are resampled to a one-kilometre resolution and projected to the Albers Equal Area projection to preserve correctness of area, to ensure that the correct number of hectares is allocated. During the calibration and validation phase, 2003 to 2010, the demand is known from the observational data; it is simply the total area of sugar cane cultivation on the Canasat map per year (Fig. 5). For 2011, the Canasat map is not available to us, but the total area is, so the demand is known. Two data sources are used to construct the demand between 2012 and 2016. From the Brazilian Land Use Model (BLUM) (ICONE, 2012; Nassar et al., 2008), an economic partial equilibrium model, preliminary results of the future development of harvested area of sugar cane in São Paulo are used. For sugar cane the harvested area is always smaller than the cultivated area, because sugar cane is a semi-perennial crop: after about six to eight harvests the cycle is interrupted and the area is renovated for a year. The harvested area from BLUM is converted to cultivated area by adding the average fraction of sugar cane fields under renewal, which is derived from Canasat data. From the Brazilian agricultural economics institute, IEA, a study is used that estimates the cultivated area of sugar cane in São Paulo up to 2016 (Torquato, 2006). As we have equal trust in both sources, the demand in the land use change CA from 2012 to 2016 is the mean of the two time series created from these sources (Fig. 5). It would be better to take into account the input uncertainty coming from the inconsistency between the two time series, but in this study we want to show how the uncertainty in the model structure and parameters propagates, without interference with the demand uncertainty.

The preferred location of the expansion of sugar cane is regulated by the total suitability map, an assembly of the no-go areas, and all suitability factors (Fig. 4). The no-go map is derived from the sugar cane zoning for the São Paulo state (Padua Junior et al., 2012). These cells are masked from the total suitability map and therefore not available for land use change. The total suitability map $\mathbf{s}_t \in [0,1]$ for sugar cane at time step t is:

$$\begin{aligned} &\mathbf{s}_t = \sum_{k=1}^{K}(w_k\cdot\mathbf{u}_{k,t}), \text{ for each } t \\ &\text{with } \sum_{k=1}^{K}(w_k) = 1 \\ &\text{and } \mathbf{u}_{k,t} = h\left(\mathbf{x}_{k,t},\mathbf{p}_{k,t}\right) \end{aligned} \tag{5}$$

In Equation (5), k is the suitability factor, with $k = 1, 2, \ldots, K$ and $w_k \in [0,1]$ is the spatially and temporally uniform weight of factor k. Furthermore, $\mathbf{u}_{k,t} \in [0,1]$ is the suitability map for suitability factor k. The function $h()$ uses the spatial attribute $\mathbf{x}_{k,t}$

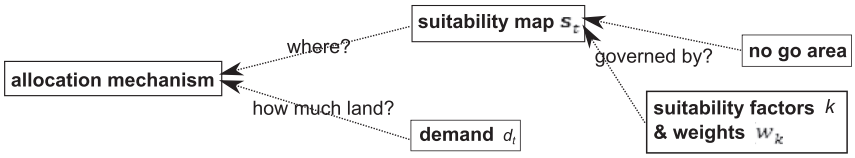

Fig. 4. Schematic representation of the land use change transition function.

and parameter $\mathbf{p}_{k,t}$ to create the proxy for land use change, and then normalizes it, i.e. linearly transforms it to a scale between 0 and 1, to obtain $\mathbf{u}_{k,t}$. The transformation is linear, because the actual shape of the relation (linear, convex, concave) between $\mathbf{u}_{k,t}$ and $\mathbf{x}_{k,t}$ is determined by the parameters $\mathbf{p}_{k,t}$ within $\mathbf{u}_{k,t}$, discussed later in this section per suitability factor k.

The model structure of the CA, $p(\mathbf{f}_t)$ (Equation (1)), is formed by the weights w_k, because they determine if a certain process, or suitability factor, is of influence on the total suitability map, and how large this influence is. The prior distribution of the weights is constructed using the following procedure, making sure that every weight covers the complete range [0,1] and that the constraint $\sum_{k=1}^{K}(w_k) = 1$ is preserved. The weight of the first (randomly chosen) suitability factor k is drawn uniformly between zero and one. The next weight is drawn between zero and one minus the sum of the previous weights. The last weight is one minus the sum of all others. In this way, all weights have the same prior probability distribution, and all weights can become high (close to one) and are also often set to (close to) zero, i.e. the process is (almost) switched off. By assimilating observations, some weights can converge to zero over time, indicating that the processes, which the associated suitability factors embody, are irrelevant in the observed system. Other factors will prove to be important. So, by determining which suitability factors are relevant, we identify the CA model structure.

The 'candidate' suitability factors and a short explanation of the processes they represent are listed in Table 1. They are referred to as candidate suitability factors, because over the course of the calibration they either prove to be relevant, by obtaining a weight above zero, or to be irrelevant, by receiving a weight of zero. The candidate factors are derived from informal discussions with experts and literature review (Lapola et al., 2010; Walter et al., 2011; Rudorff et al., 2010; Macedo and Seabra, 2008; Sparovek et al., 2007, 2012; de Souza Soler and Verburg, 2010; Aguiar et al., 2011). Sugar cane in the neighbourhood ($k = 1$) is expected to be important because larger plantations usually require less money per hectare as equipment and infrastructure can be shared (economies of scale). Also, a group of existing sugar cane fields usually already has a mill, in which the sugar cane is crushed, in the vicinity, so the sugar cane from new fields in the neighbourhood could go to the same mill. Travel time, and thereby transportation costs, to São Paulo city ($k = 2$) could be of influence because the ethanol is distributed through there, so São Paulo city is the main market. It is assumed that transportation occurs by truck only (Macedo and Seabra, 2008). Potential yield ($k = 3$) is important for the potential profits per hectare. We use a potential yield map created from physical landscape properties and climate data by the IIASA (Tóth et al., 2012). Slope ($k = 4$) is critical, because the São Paulo state tries to eliminate pre-harvest burning, with its negative impacts on human health and on the environment due to the emission of pollutant gases (Aguiar et al., 2011). Pre-harvest burning can be banned when manual harvesting is replaced by mechanical harvesting, which does not require burning. However, the harvest machines cannot operate on sloping ground. The state law that promotes sustainable production practices for sugarcane in São Paulo State therefore induces new sugar cane agrarians to avoid slopes above 12%. Random noise ($k = 5$) is used as a suitability factor to include local allocation choices that cannot be captured by a specific process or attribute at this model scale.

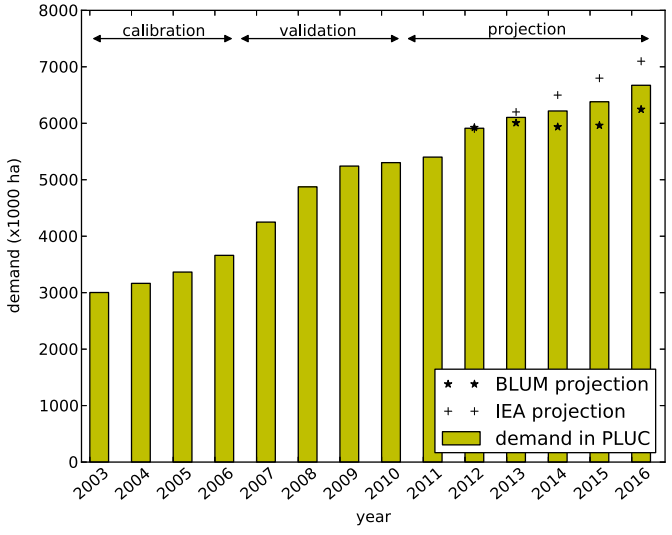

Fig. 5. Demand for sugar cane area from 2003 to 2016. Demand in the calibration and validation phases comes from the Canasat maps. Demand in the projection phase is the mean of two projection time series from BLUM and IEA.

The tuning of the parameters of the model, or calibration, relates to finding the posterior distribution of all parameters, $p(\mathbf{p}_t)$, within the suitability factors. In total, there are five parameters to calibrate: $\mathbf{p}_t = [f, l, a_2, a_3, a_4]$, which are explained in the following. Suitability factor 1, the neighbourhood effect, is defined as:

$$\mathbf{u}_{1,t} = norm\left(-\mathbf{x}_{1,t}^2 + 2 \cdot f \cdot \left(\frac{l}{c}\right)^2 \cdot \mathbf{x}_{1,t}\right), \quad \text{for each } t \tag{6}$$

In Equation (6), the number of neighbours being sugar cane in the neighbourhood window in a certain time step is $\mathbf{x}_{1,t}$. Parameter c is the cell length, in this study fixed at 1000 m, and parameter l (m) is the window length of the window that determines whether or not a cell belongs to the neighbourhood. And f is the 'preferred' fraction of neighbours being sugar cane of the total number of neighbours, $(l/c)^2$, within the window. The function $norm()$ normalizes its contents. The prior distribution of l is lognormal, $l = e^Z l$, with $Z_l \sim N(8.5,0.7)$, which results in a median of around 5000 m and a mean of 3000 m. For these values, the window, $(l/c)^2$, is the extended and direct Moore neighbourhood, two of the most commonly used neighbourhood types. The prior distribution of f is uniform, $f = Z_f$, with $Z_f \sim U(0,1)$. If, for example, f equals 0.5, the highest suitability ($\mathbf{u}_{1,t} = 1$) occurs where half of the neighbours in the window is sugar cane. A reason why it could be the case that the neighbourhood shows 'gaps' without sugar cane is that a law, the Forest Code, requires that a certain portion of private farmland is set aside for natural vegetation preservation. In São Paulo, being outside the Legal Amazon Region, this is 20% (Sparovek et al., 2012). Another reason why more sugar cane in the neighbourhood is not always better is that the mill in which the sugar cane is crushed, has a maximum crushing capacity per season. In São Paulo the average maximum capacity is $1.9 \cdot 10^6$ tonnes (Walter et al., 2011). When this capacity is reached, it is not necessarily an economic advantage anymore to create new sugar cane fields close to existing ones, as a new mill has to be built anyway.

Suitability factors 2, 3 and 4 from Table 1 are calculated as:

$$\begin{aligned} \mathbf{u}_k &= norm\left(1 - \mathbf{x}_k^{a_k}\right), \quad \text{for } k = 2,4 \\ \mathbf{u}_k &= norm\left(\mathbf{x}_k^{a_k}\right), \quad \text{for } k = 3 \end{aligned} \tag{7}$$

In Equation (7), $\mathbf{x}_k$ is the attribute of suitability factor k. In the case of attributes travel time to São Paulo ($k = 2$) and slope ($k = 4$), lower attribute values lead to a higher suitability, while for potential yield ($k = 3$) higher attribute values lead to a higher suitability. The parameter a_k determines the shape of the suitability function. A value of one results in a linear function, meaning that suitability increases or decreases linearly with the increase in the value of the attribute. For $0 < a_k < 1$, the shape of $\mathbf{u}_k$ is concave, and for $a_k > 1$, the shape is convex. The prior distribution of

Table 1
Candidate suitability factors for sugar cane in São Paulo.

k	Suitability factor	Process represented
1	Sugar cane in neighbourhood	Economies of scale
2	Travel time to São Paulo[a]	Transportation costs to the main market
3	Potential yield	Profits
4	Slope	Mechanization potential
5	Random noise	Local allocation choices (unexplained)

[a] Calculated as distance divided by speed. Speeds on different road types are taken from de Souza Soler and Verburg (2010).

a_k is lognormal, $a_k = e^{Z_{a_k}}$, with $Z_{a_k} \sim N(0, 1.8)$, which results in a distribution of a_k with a median of one, i.e. a linear relation between $\mathbf{u}_k$ and $\mathbf{x}_k$. The uncertainty of the attributes, $\mathbf{x}_k$, is based on information provided with the datasets (Tóth et al., 2012; Jarvis et al., 2008). Only the locations of roads and São Paulo city, used for suitability factor 2, are assumed to be known, and therefore used deterministically. Note that these suitability factors and their uncertainty remain static over time. In reality they change, e.g., new roads can be build and potential yield can decline due to land degradation, but these processes are not taken into account due to a lack of data.

Suitability factor, 5, has no parameters. Its suitability is equal to its attribute, which is a uniformly distributed random spatial field, varying between zero and one, drawn separately in each time step, so $\mathbf{u}_{5,t} = \mathbf{x}_{5,t}$.

2.3.3. Spatial metrics

The purpose of land use change models is usually not, and should not be, to simulate precisely the land use of each single cell in each year (Parker et al., 2008). More realistic is to try to capture certain spatio-temporal patterns. Therefore aggregated measures or spatial metrics are often more useful for calibration than location-based methods (Pijanowski et al., 2006). To correctly identify the system dynamics, it is essential to evaluate multiple system characteristics. So, multiple spatial metrics should be assessed in the model structure identification and calibration, observed at different system levels (Grimm and Railsback, 2012). Three spatial metrics were selected based on their complementarity (global vs. regional, configuration vs. composition (Csillag and Boots, 2005)). 1) The fraction of sugar cane in 150 × 150 km blocks. This metric ensures that the demand for sugar cane in the São Paulo state is distributed in correct proportions over regional areas. The metric is regional and based on composition. 2) The total number of interconnected sugar cane patches. This signifies whether all sugar cane cropland is connected into one patch or distributed over many patches. 3) The landscape shape index, q_t, calculated as (Pijanowski et al., 2002):

$$q_t = e_t / \min(e_t) \tag{8}$$

Herein, e_t (m) is the total length of the edge of the sugar cane patches in time step t, min (e_t) (m) is the minimum total length of edge for a maximally aggregated sugar cane patch, attained if all sugar cane is grouped into one square patch. So, it indicates the shape of the patches, e.g., very compact or more crooked. The last two metrics are global and based on configuration, so that a balance between global and regional, and composition and configuration is obtained in the model calibration and validation.

Validation is done using the same spatial metrics used for calibration, as the performance criteria should match the model purpose and thus the calibration criteria (Rykiel, 1996). The root mean square error (RMSE) and 95% confidence intervals are used to quantitatively compare the spatial metrics obtained from the model projections and from the observational data of the same period.

2.3.4. Observational data

The observational data are eight annual maps of sugar cane occurrence, classified from Landsat images by INPE for the Canasat project (Rudorff et al., 2010), with a resolution of 30 m and a temporal extent from 2003 to 2010. The data are resampled to a 1 km resolution.

In order to solve Equation (4) (particle filter), not only the mean observations, $\mathbf{o}_t$, i.e. the observed spatial metrics, have to be known, but also their error covariance, $\mathbf{R}_t$. As the error of the Canasat maps is not known, at least not for each cell, the error is determined by generating possible realizations of these maps. Hereto we use a stochastic simulation procedure that is widely applied to predict the uncertainty of an attribute at unknown locations, given a set of known locations (Pebesma and Wesseling, 1998). We use it to assess the uncertainty of an attribute at locations for which we already know the attribute value. Herein the following actions are taken for each time step in the calibration period:

1. Experimental semi-variances (Cressie, 1993) are calculated and plotted. This is done based on a Boolean map, in which sugar cane is one and no sugar cane is zero.
2. A semi-variogram model (Cressie, 1993) is fitted using the software gstat (Pebesma, 2004). An exponential model was chosen, because it yielded the best fit and because this model is usually a good choice when several patterns interfere (Burrough and McDonnell, 1998), which can be expected for land use patterns that are often governed by many drivers of location.
3. Cross-validation (Burrough and McDonnell, 1998) is performed to check the semi-variogram model.
4. Gaussian simulation (Pebesma and Wesseling, 1998) is used to create 100 potential spatial fields of scalar values, in which values close to zero indicate that there is probably no sugar cane and values close to one that there probably is.
5. A threshold is applied to these 100 fields to turn them into Boolean maps again. To include not only configuration errors (sugar cane located in the wrong cell), but also composition errors (the total area of sugar cane in the map is wrong), this threshold is a normally distributed stochastic parameter with a mean of 0.5 and a standard deviation of 0.1. As a result, some of the realizations will have a larger sugar cane cropland coverage than others.
6. From each of the 100 realizations, the three spatial metrics are derived.
7. The covariance matrix $\mathbf{R}_t$ is calculated from these spatial metrics.

2.3.5. Scenarios

To summarize, the land use change model (Fig. 4) is run for the case study. During run time Equation (3) is applied at each filter moment to find $p(\mathbf{z}_t^i|\mathbf{o}_t)$, the posterior probability of land use change model particle i. The system state $\mathbf{z}_t^i$ and the observations $\mathbf{o}_t$ are compared in the form of the three spatial metrics (Section 2.3.3), derived from respectively the model output and the Canasat maps. Because $p(\mathbf{z}_t^i|\mathbf{o}_t)$ contains, besides the system state, also the transition rules $p(\boldsymbol{f}_t)$, where $\boldsymbol{f}_t \leftarrow w_k$ (Equation (5)) and the parameters $p(\mathbf{p}_t)$, where $\mathbf{p}_t = [f, l, a_2, a_3, a_4]$ (Equations (6) and (7)), these are updated as well when the ensemble of model runs is updated using sequential importance resampling.

As one should not use the same set of data for calibration and validation, we use a split-sample approach over space and time in all scenarios. Two years (2004–2005) are used for calibration and five (2006–2010) for validation. More years are used for validation than for calibration to be able to study to what extent the performance decreases over time. In addition, the data are split over space: half of the ten 150 × 150 km blocks is used for calibration (from this point onwards called calibration blocks), and the other half is not (from this point onwards called validation blocks). In this way, the performance of the calibration blocks and the validation blocks can be evaluated separately, to assess to what extent the model can be used for areas where no calibration data are available. Then, we use the calibrated and validated model for land use change projections up to 2016.

Four calibration scenarios were designed (Table 2). The first is the reference case, i.e. what would happen without filtering, thus only Monte Carlo simulation. The second scenario is meant to show the effect of the particle filter and to evaluate the two different sets of five blocks (sample split over space), as explained above. The third scenario uses fewer blocks, i.e. a smaller spatial coverage, to test the influence of observational data availability, as time series of good quality observations are often not obtainable (Straatman et al., 2004). In reality, this could represent a situation in which one has incomplete data for calibration, e.g., remote sensing images partly covered by clouds. The fourth scenario is meant to discuss the matter of model structure identification. In many calibration efforts the model rules are preliminarily set and only the parameters are calibrated. This scenario represents that situation in order to evaluate the difference between pre-set (this scenario) and stochastic model rules (scenario 2). All scenarios are run using $N = 25{,}000$ particles.

3. Results and discussion

3.1. Realizations of observations

For the filter moments, realizations of the observations were created, using Gaussian simulation as explained in Section 2.3.4. These realizations represent potential instances of sugar cane cropland maps and are used to determine the covariance matrix, $\mathbf{R}_t$. Fig. 6 shows a close up of nine of the hundred realizations for 2005. On the one hand one can identify the configuration errors (sugar cane cropland located in the wrong cell). For example, variations in the shape of the large patch in the middle of the close up. On the other hand, the effects of composition errors (the total area of sugar cane cropland in the map is wrong) are visible. The upper left realization, for instance, shows a larger total sugar cane cropland area than the one below it.

Table 2
Calibration and validation scenarios.

#	Purpose	Model structure (weights)	Parameters	Filter in t =	Number of blocks
1	Reference case	Stochastic	Stochastic	–	–
2	Filtering	Stochastic	Stochastic	2004, 2005	5
3	Less observational data	Stochastic	Stochastic	2004, 2005	2
4	Model rules preliminarily set	Deterministic	Stochastic	2004, 2005	5

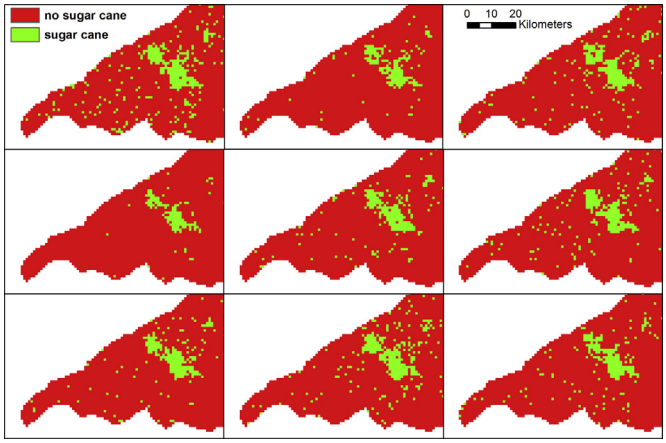

Fig. 6. Close up of the most Western wedge of São Paulo of nine realizations of the observations from 2005.

3.2. Model structure identification

The evolution of the model structure $p(\mathbf{f}_t)$ is illustrated by the evolution of the weights w_k of the five candidate suitability factors over time for scenario 2 (Fig. 7). All weights have the same prior lognormally shaped distribution between zero and one. Over time, some particles are filtered out and others are copied. After the first filter moment (2004), of which the results can be seen in 2005 (Fig. 7), the effects are small. But, for example, for the random noise suitability factor, the weight distribution is narrowed: a land use change model structure in which random noise is the only relevant suitability factor is rejected. Also, for potential yield and travel time to São Paulo high weights have become less prevalent in the ensemble. In the second filter moment (2005), the distributions converge further, e.g., the weight of slope converges towards high values; its posterior distribution has two peaks around 0.9 and 0.5. This means that slope is a very important factor in the allocation of sugar cane cropland. Slope was expected to be important, because the São Paulo state tries to eliminate pre-harvest burning, as explained in Section 2.3.2. We had, however, not foreseen that slope is so much more important than the other factors. Travel time to São Paulo city, for example, was expected to be more important. The results that it is not (it has a weight distribution between 0.0 and 0.2), could be because the whole study area is relatively close to São Paulo city. Another explanation can be the fact that the sugar cane is not transported to the city directly. First the sugar cane is transported to sugar cane mills, where it is converted to ethanol. The ethanol in its turn, is usually transported to São Paulo city. Ethanol, though, is a much higher value and higher energy density product, for which transportation costs, and therefore travel times, are less important. We have considered to include travel time to the sugar cane mills as a suitability factor, but the problem with this is that new mills

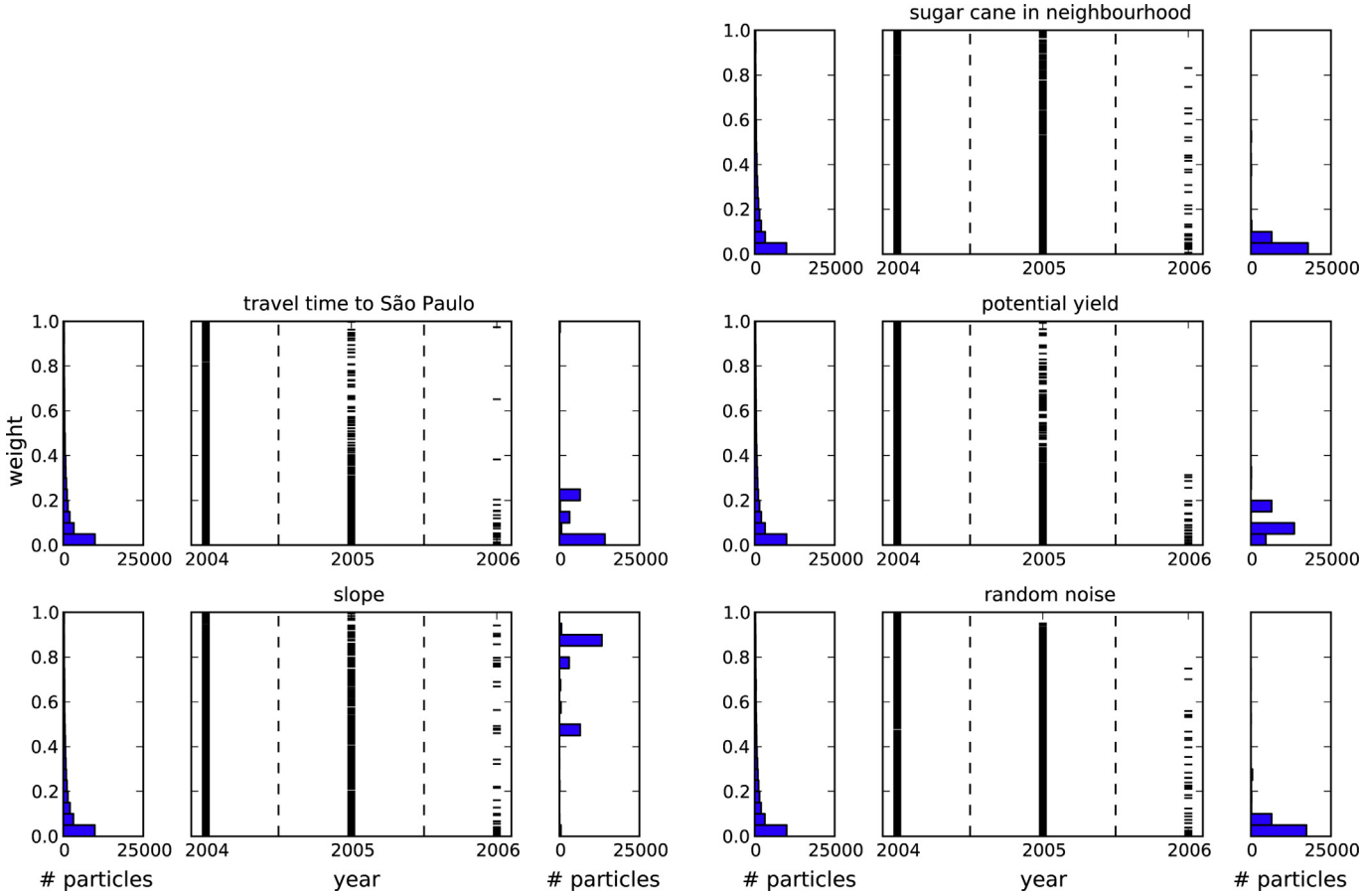

Fig. 7. Evolution of the weights w_k of the candidate suitability factors over time for scenario 2, with filtering in 2004 and 2005. For each suitability factor the black horizontal lines in the centre panel are a random selection of 10% (for visualization purposes) of the particles, the blue bars on the left represent the prior distribution, and the blue bars on the right represent the posterior distribution of the weight (see also diagram structure on the top left).

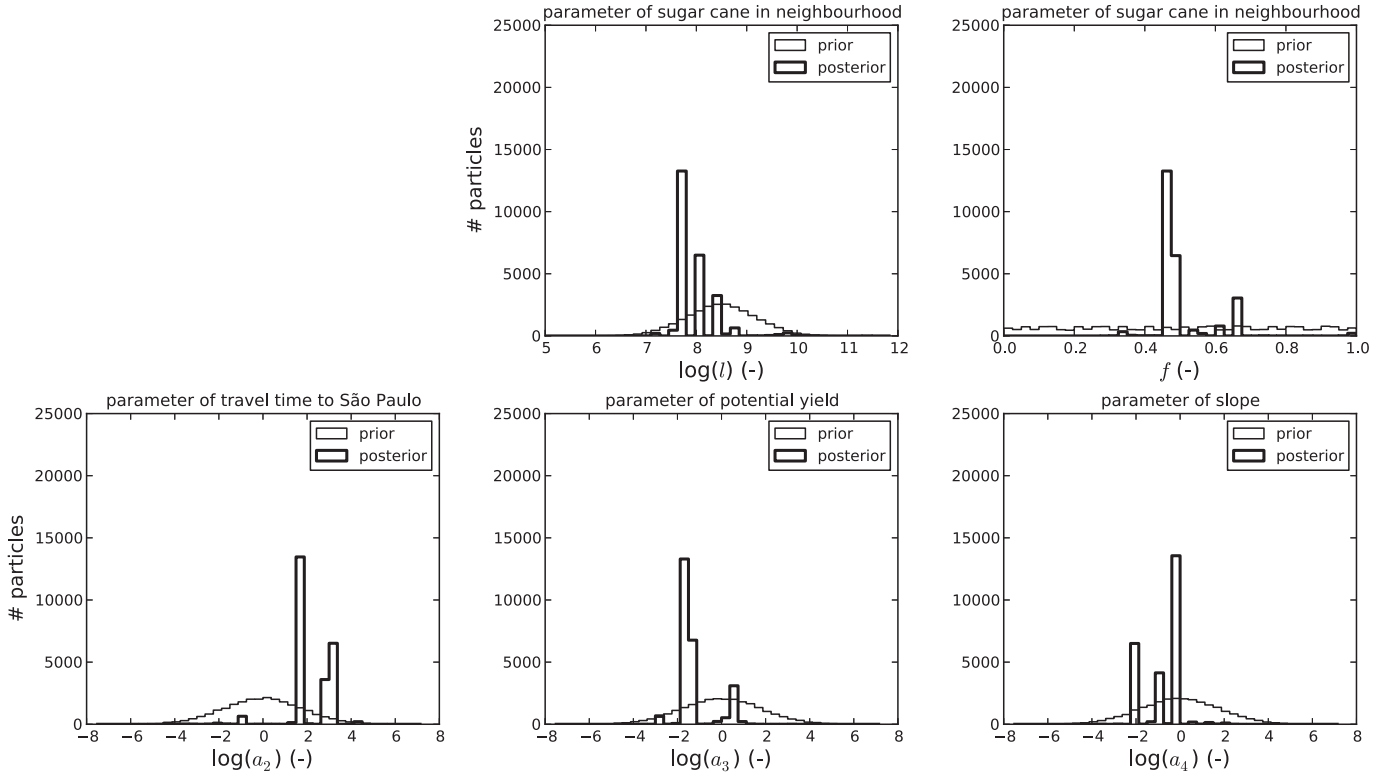

Fig. 8. Prior (thin line) and posterior (thick line) distributions of all parameters for scenario 2.

emerge quite often, and since we do not know where, forecasting becomes problematic with this model structure. An additional reason for relatively low importance of travel time could be that an ethanol pipeline is present running through the high-density sugar cane cropland area in the middle north part of São Paulo state to São Paulo city. More pipelines are planned in the future. Macedo and Saebra (2008) expect that by 2020 20% of the ethanol in Brazil will be transported through pipelines.

Some weights converge to values close to zero, as is the case for sugar cane in the neighbourhood and random noise, which means the processes are less relevant. The fact that random noise obtains a low weight in the posterior is a good sign. It indicates that unexplained local allocation choices have little influence on the regional and global sugar cane pattern. However, it does not converge to zero entirely, so it is not completely irrelevant.

3.3. Calibration

At the same time the parameters within the suitability factors have been calibrated (Fig. 8). The prior of the logarithm of window length *l*, of the sugar cane in neighbourhood suitability factor, is normally distributed and the posterior has converged to values around 8. This means that the best neighbourhood is the direct Moore neighbourhood, as $e^8 \approx 3000$ m, which is 3 cell lengths. The prior distribution of the preferred fraction of neighbours being sugar cane *f* is uniformly distributed between zero and one. The posterior distribution has two peaks. The largest lies around 0.5, indicating that the preferred number of neighbours being sugar cane is half of the total number of neighbours in the neighbourhood. Because the posterior distribution of window length *l* has a value of about 3 cells, this means that the highest suitability for the neighbourhood function is reached when four out of the eight direct neighbours are sugar cane. This could imply that farmers set aside part of their farmland to meet the terms of the Forest Code. But to verify this, a further study on the land use of the non-sugarcane cells in the window should be done. When these cells are indeed natural surroundings, it could be the case, but when this land is used otherwise, e.g., for other crops or grazing of livestock, a different reason exists for the scattered pattern. At the moment we are unable to examine this, because we have no map of the other land uses in the São Paulo state that is sufficiently detailed in space, time and attributes.

The lower three panels in Fig. 8 represent the logarithm of the parameter a_k, which determines the shape of the suitability functions of travel time to São Paulo ($k = 2$), potential yield ($k = 3$) and slope ($k = 4$). The prior of the natural log of all three parameters is normally distributed, with a median of zero, so that the median of $a_k = e^0 = 1$, which results in a linear relationship between the attribute value and the suitability. The posterior distribution of a_2 for travel time to São Paulo has its peaks above zero. Applying Equation (7), we see that the resulting shape of $\mathbf{u}_2$ plotted against $\mathbf{x}_2$ is convex (Fig. 9). This means that up to a travel time of about ten hours from São Paulo city the suitability is high and almost constant, and further away it quickly drops to zero. On the map this shift arises far away from main roads in the North-western and South-western parts of the study area only. From that point onwards transportation costs possibly become too high.[1]

[1] If costs versus revenues are truly the reason for the position of the tipping point of a_2, it should be noted that this point is very sensitive to e.g., changes in fuel prices. The model could be improved by calculating transportation costs instead of travel time, so that these economic variables are reflected in the suitability factor. However, obtaining data to accurately do that is time-consuming, as e.g., regulations, taxes and fuel prices can vary widely per administrative unit in Brazil.

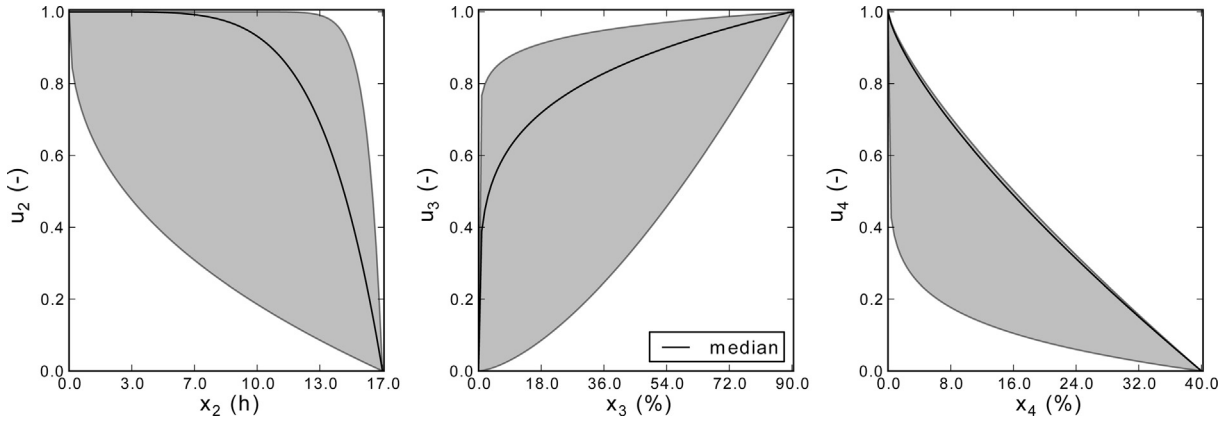

Fig. 9. Suitability distributions $\mathbf{u}_k$ for the attributes $\mathbf{x}_k$: travel time to São Paulo ($k = 2$), potential yield ($k = 3$), and slope ($k = 4$), given the median (solid black line) and the 95% confidence interval (grey area) of the posterior distribution of a_k.

The posterior distribution of a_3 for potential yield has its largest peak around −2. Applying Equation (7), we see that $\mathbf{u}_3$ plotted against $\mathbf{x}_3$ has a convex shape (Fig. 9). The curve shows that even soils with relatively low yield, in the region of 15% of the maximum attainable yield, are suitable for sugar cane cultivation. This is in line with information provided to us by experts from CTBE (Brazilian Bioethanol Science and Technology Laboratory), who stated that in the São Paulo state all soils are good enough to cultivate sugar cane and the soil must be prepared anyway, so the precise quality is not that important. The climate is also uniformly good; in the entire state sugar cane can be cultivated without irrigation.

The posterior distribution of slope has its largest peak at approximately zero. Therefore, the median of $\mathbf{u}_4$ plotted against $\mathbf{x}_4$ has a linear shape (Fig. 9), meaning that the suitability decreases linearly with the increase of slope. It should be noted that slopes higher than 12% are included in the no-go area. Therefore, in the model sugar cane cannot be allocated on high slopes anyway, so the part of the graph where $x_4 > 12\%$ has no effect in the CA.

In scenario 3, in which less observational data are assimilated, the parameters converge to similar values, but the distributions remain much broader. This is because less information is available on whether a certain parameterization performs well. In scenario 4, parameter distributions are narrower again. They have converged to different values than scenario 2 and 3, to correct for the fixed model structure, which is different from the optimal model structure in scenario 2 (Fig. 7). For example, window length l, obtains a median of about 13 km. Apparently, the fact that sugar cane is the neighbourhood was given a too high weight, can be partly compensated by calculating the number of neighbours in a larger window.

3.4. Validation

Fig. 10 compares the root mean square error (RMSE) of the validation time steps 4 to 8 (2006–2010) of all scenarios for all three spatial metrics for the calibration and the validation blocks separately. The RMSE is a frequently used measure of the differences between a modelled and an observed variable. As the measure is scale dependent (Hyndman and Koehler, 2006), it has a relative meaning only, so we merely use it to compare between scenarios. For absolute comparison, the modelled spatial metrics and the observed spatial metrics are compared for scenarios 1, 2, and 4 in Fig. 11.

In general, it can be observed that in the two scenarios in which the particle filter was applied, the width of the 95% confidence interval is significantly smaller, and the relative difference in width between the scenarios with and the scenario without filtering increases over time. For example, the width of the confidence interval in scenario 2 is less than 10% of the width in scenario 1 in 2016 (Fig. 11). The scenarios in which the particle filter was applied have a lower RMSE (Fig. 10). This implies that the particle filter reduces the uncertainty in the ensemble of model runs in a way that brings the output values closer to the observed values. For the calibration blocks, the general trends are similar to the validation blocks. Overall, the performance is a bit better in calibration blocks than in the validation blocks, as expected.

Scenario 2 outperforms the other scenarios regarding all metrics throughout the modelled period, except for two years (Fig. 10). Concerning the fraction sugar cane per block, the median of the model output and the observations differ less than 1% up to 2006, and for three out of the five blocks even all the way up to 2010 (Fig. 11). In scenario 1, the divergence of the two medians starts already in 2004 and concerns four instead of two blocks. This shows that the particle filter has managed to resample the ensemble of particles in such a way that the sugar cane expansion behaviour is corrected in the two other blocks, however not all the way up to 2010. The median of the model output and the observations for landscape shape index differ less than 1% up to 2006 for both scenario 1 and 2. After that, the modelled landscape shape index is too low (maximum of 10%), and the observations do not fall within the 95% confidence interval in scenario 2. So, the particle filter does narrow the confidence interval and reduce the RSME over the complete validation period, but further away from the filter moments the ensemble is not successful anymore in projecting either the fraction of sugar cane for two out of the five blocks or the landscape shape index. This can be a result of either incorrect model identification, or non-stationarity in the land use system itself. The implication of this is that, if one wants to use the identified CA to assess an impact of land use change that uses these blocks or the landscape shape index as a criterion, the reliability of this impact analysis is low. Yet, the fact that there is little difference in performance between the calibration and validation blocks for scenario 2 (Fig. 10) indicates that the

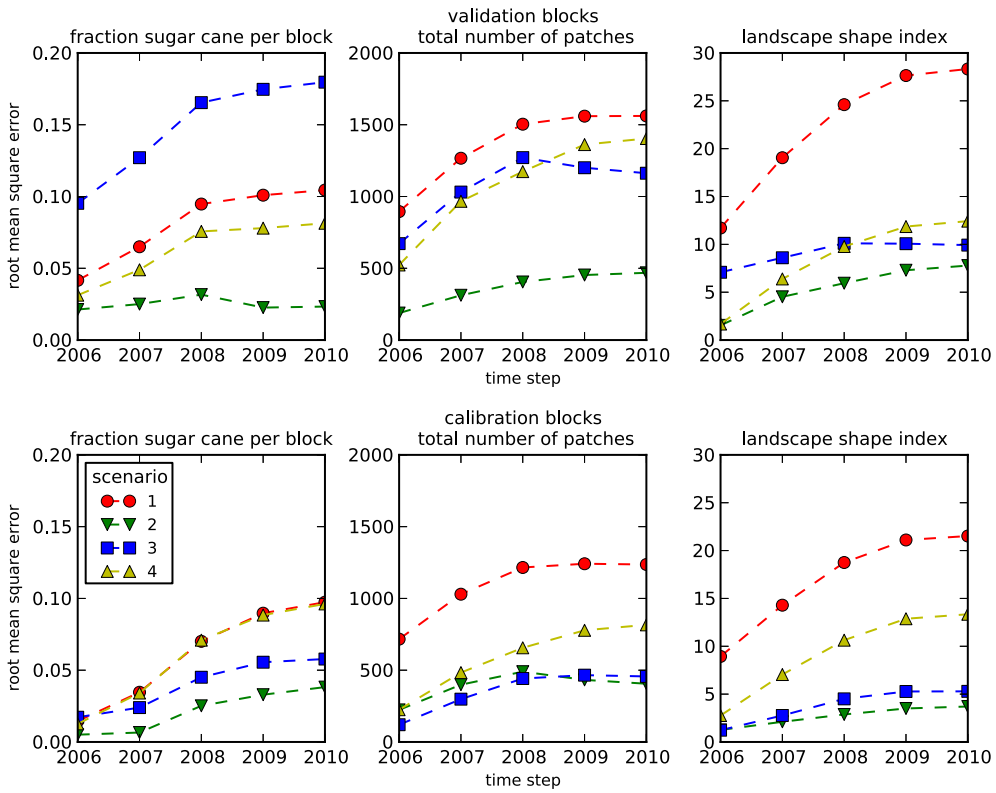

Fig. 10. Root mean square error for all spatial metrics for all scenarios for the validation blocks (top) and calibration blocks (bottom) for scenario 1 (reference case), 2 (particle filter), scenario 3 (less observational data), and 4 (preliminarily set model rules).

identified model structure and parameters perform equally well in the part of the study area for which no observations were assimilated. With information from half of the study area, the particle filter is able to identify model structure and parameters resulting in the same performance in the other half of the area.

The effect of less observational data (scenario 3) is largest for the fraction of sugar cane per block (Fig. 10), where it performs even worse than the reference scenario (scenario 1). The RMSE is more than four times as large as for scenario 2. For the calibration blocks scenario 3 seems to perform better, but is calculated only for the two calibration blocks instead of five as is the case for the other scenarios. Because of this it is problematic to compare the RMSE of scenario 3 with the other ones for the calibration blocks.

For scenario 4 (pre-set model structure but calibrated parameters), the decrease in performance is especially large for the number of patches (Fig. 10). Fig. 11 shows that this scenario is completely unable to capture the trend in this metric. It predicts continuous decrease from the start, while there should be an increase.

To further compare the predictive power of the reference and the particle filter scenario, Fig. 12 shows the maps of the probability of sugar cane cropland coverage in 2007 and 2016 for both scenarios the observational data in 2007. In 2007, two filter moments have passed in scenario 2, so its land use map differs a lot from the one of the reference scenario. In scenario 1, sugar cane cropland has expanded along edges of existing patches. The area in the Southern part of São Paulo state has probabilities of zero because this is defined as the no-go area for sugar cane (Padua Junior et al., 2012). Furthermore, almost all cells have a, although low, probability to be occupied by sugar cane in 2007. In the output of scenario 2, the uncertainty of where sugar cane will be located has been greatly reduced; most cells are either red (probability of zero) or green (probability of one). The majority of the expansion has taken place in the North-West. Both the location and the configuration (scattered) of the expansion in this scenario result resemble the observations much better than scenario 1.

For 2016 the same differences between scenario 1 and 2 appear; scenario 1 is much more uncertain, although the uncertainty in scenario 2 has also increased, as the time period since the last filter moment is long. Scenario 2 indicates the highest probability of expansion mainly in the western part of São Paulo state. Obviously, there are no observations for 2016, so validity cannot be checked. Yet, the sugar cane cropland expansion sites up to 2020, given by Lapola et al. (2010, p. 2) resemble the locations with a high probability in our results closely.

To ensure that the model structure and parameters obtained with the particle filter are not dependent on partition of the blocks for calibration and validation, scenario 2 was repeated with a different, again randomly drawn, block division. In Figs A1, A2, and A3 in the Appendix the equivalents of respectively Figs. 7, 8, and 11 are shown for the new sets of calibration and validation blocks. The results are deemed comparable, so we reject the possibility that the results shown in this section are a product of the partition of the area into calibration and validation blocks.

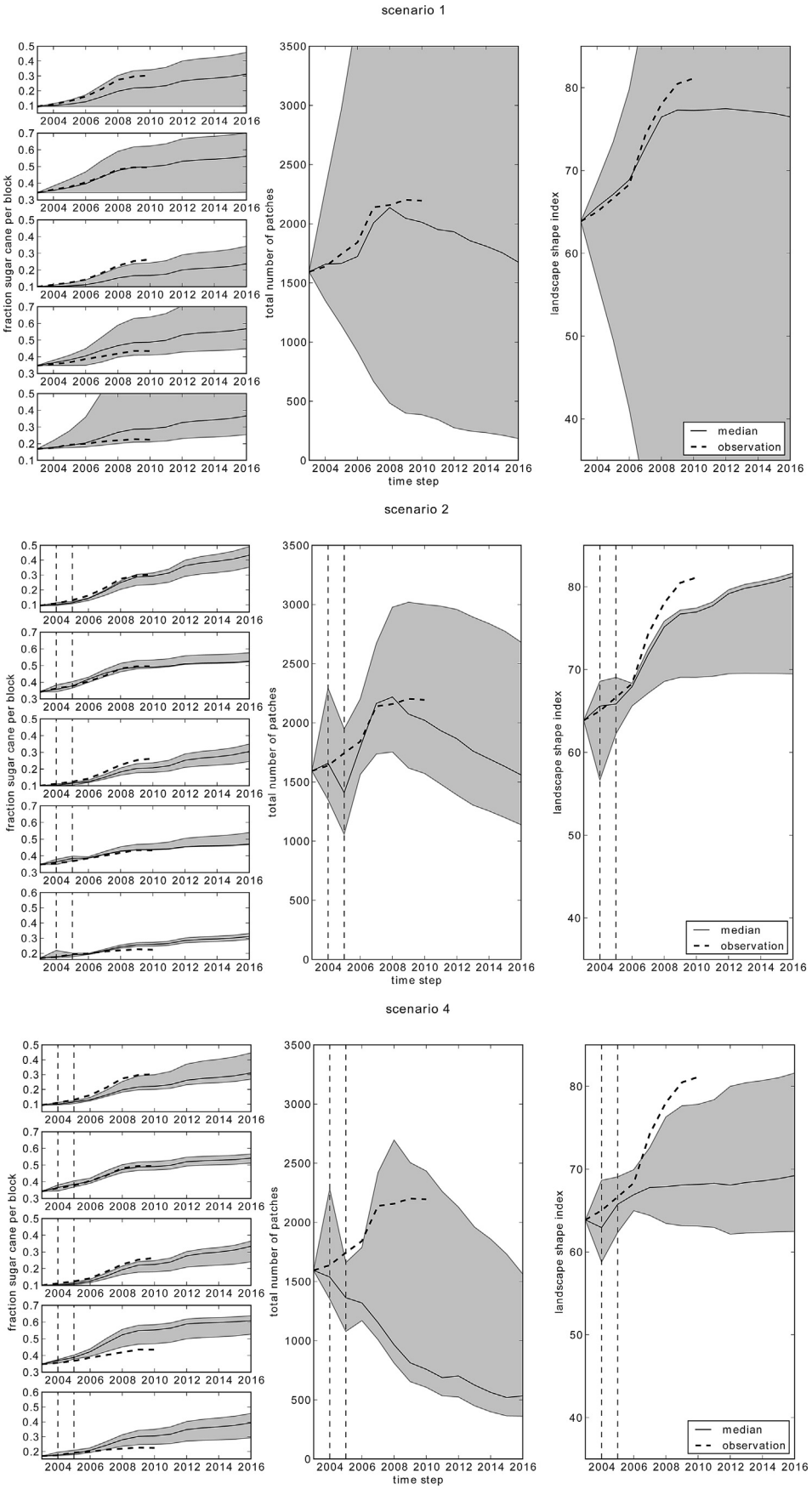

Fig. 11. Comparison of modelled median (solid line), 95% confidence interval (grey area) and observations (dashed line) of the spatial metrics for the five validation blocks for scenario 1 (reference case), 2 (particle filter) and 4 (preliminarily set model rules). The filter moments are indicated with vertical dashed lines.

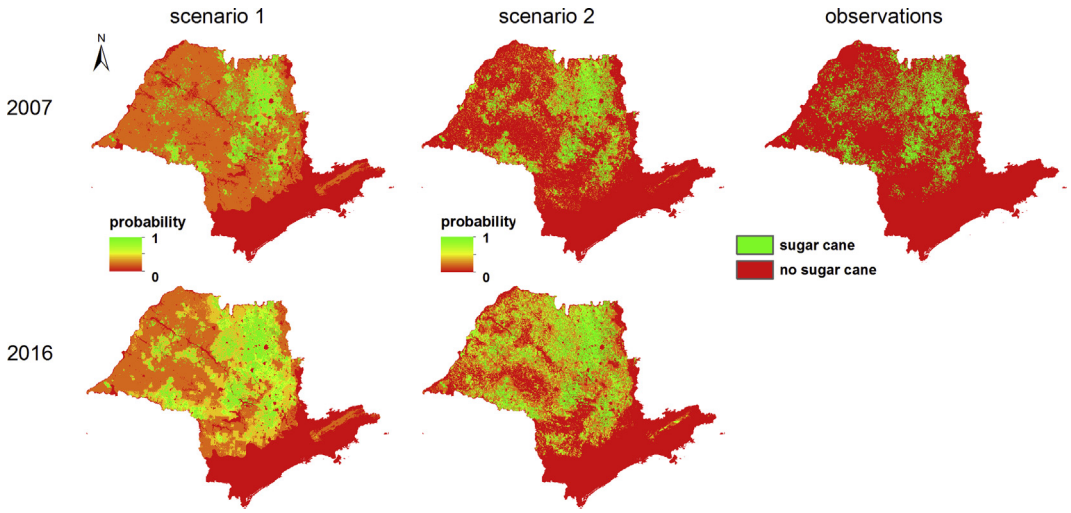

Fig. 12. The probability of sugar cane cultivation for scenario 1 (left) and scenario 2 (centre), and the observations (right), for 2007 and 2016.

4. Conclusion

The method used here simultaneously identifies the model structure and calibrates the parameters of a land use change cellular automaton (CA) by sequential assimilation of observations, using a particle filter. The method uses the (subjective) knowledge of experts to define which processes or drivers might be important in the system, and applies the (objective) information from observations to adjust the model structure and calibrate the parameters. With the candidate suitability factors chosen in this study and the observational data used, the particle filter identified a probability distribution of model structures and associated parameters that could forecast the chosen spatial metrics fairly good. Nevertheless, performance clearly decreased over time, further away from the filter moments (Figs. 10 and 11). So, using this calibration technique, information about the land use system is gained and short-term land use projections are clearly improved, but projections of more than a few years ahead are not very reliable. It remains, however, a question whether this is a deficit of the calibration technique, or a result of non-stationarity in the land use system itself. The assessment of the persistence of land use change drivers and possible changes in their relative importance over time, is the next stage of our research.

In areas where no observational data were assimilated (validation blocks), the model performed about as well as in the areas for which observations were assimilated (calibration blocks) (Fig. 10). So, spatially incomplete datasets, regional land survey data, or clouded remote sensing images can still provide valuable information for this CA identification. Also, data gaps in time are not a problem, as the sequence of filter moments does not need to be continuous, i.e. there may be gaps in the time series of observations that is assimilated. These two characteristics of the used method are a considerable advantage given the fact that time series of good quality land use maps are rare (Straatman et al., 2004). Only when the area of observational data availability became significantly smaller, the model performed a lot worse in the unknown areas (validation blocks). Besides the improvements on the land use change model itself, the technique also improves land use system knowledge, because the result of the model structure identification provides information on the relative importance of the drivers. For example, in our case study slope turned out to be much more important for sugar cane cropland allocation than expected in advance.

It is shown that the particle filter method can be used not only to calibrate the parameters inside the transition rules, but also to find the relative importance of the transition rules, the model structure. The importance of taking into account the identification of the model structure is shown by running a scenario with a pre-defined model structure. As expected, it performed worse compared to the scenario in which the model structure was identified by the particle filter. However, the choice for this specific, pre-defined structure was of course arbitrary, so this result should be interpreted with caution. A 'real' expert might have chosen a completely different model structure, possibly performing better.

Before carrying out the method presented in this paper, one should also contemplate the aim of the CA modelling effort. The calibration target, i.e. the spatial pattern that the model should be able to reproduce, should match the modelling aim. If, for instance, the effect of future land use change on animal passageways is studied, connectivity of patches is probably an important characteristic. Hence, one or more measures of connectivity should be used as a calibration and validation target. In this study, three spatial metrics were used. It turned to be complicated to correctly reproduce the landscape shape index in the longer run. This should be taken into account when studying impacts that rely on these properties.

Although it is an advantage that the search space, i.e. the prior distributions of parameters and model structure, can be defined by experts, it should be kept in mind that this prior information has a large effect on the outcomes. The selection of candidate suitability factors and the prior distribution of their weights should be performed carefully. The potential solution to incorporate a huge number of candidate suitability factors to make sure that all possible drivers are considered is not feasible, because the addition of one parameter makes the number of required particles increase exponentially (Bengtsson et al., 2008). Many CAs have such a large number of parameters that computation time and disk space become severe constraints. Three possible solutions, that can also be combined, are: 1) to fix parameters having little influence on outcomes or having evident values and to calibrate only the remaining ones, 2) to apply a more advanced particle filter scheme giving similar results with a lower number of particles thus reducing the required run time

and disk space (Spiller et al., 2008; Jeremiah et al., 2012), and 3) to use super computers or cluster machines, to allow for a larger number of particles in the data assimilation scheme, thereby enabling an increase in the number of parameters that can be calibrated.

An advantage of the method shown is that model uncertainty and observation uncertainty are taken into account. It must be acknowledged that the uncertainty in the observational data is constructed by making realizations using Gaussian simulation. So, the assumed observation uncertainty can be incorrect, but at least observation uncertainty is not ignored, as it is in many other predictive studies (Ivanovic and Freer, 2009). This means that our output uncertainty encompasses errors from model structure, parameters and observation (calibration) data. Such a full scope error propagation assessment is, to our knowledge, new in land use change CA modelling. In our opinion, it can be very useful, for example, to determine for which future time frame the results are reliable enough to base decisions and policies on. A practical application of how uncertainty information can be used in decision making is given by, e.g., Aerts et al. (2003) and Verstegen et al. (2012).

Acknowledgements

This work was carried out within the BE-Basic R&D Program, which was granted a FES subsidy from the Dutch Ministry of Economic affairs, agriculture and innovation (EL&I). We are grateful to the National Institute for Space Research in Brazil (INPE) for providing the Canasat maps that were used as observational data.

Appendix A

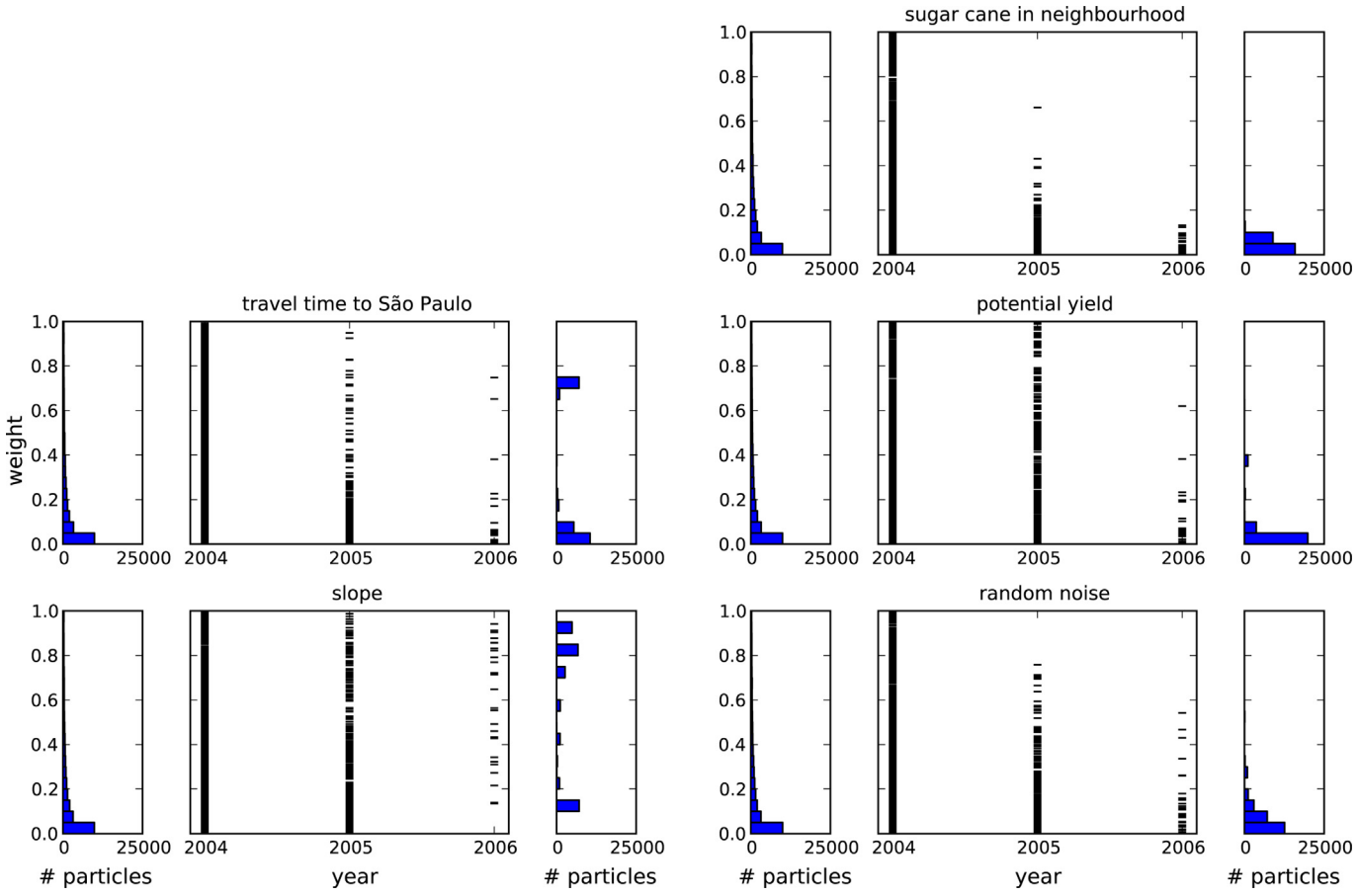

Fig. A1. Evolution of the weights w_k of the candidate suitability factors over time for the new block division, with filtering in 2004 and 2005. For each suitability factor the black horizontal lines in the centre panel are a random selection of 10% of the particles, the blue bars on the left represent the prior distribution, and the blue bars on the right represent the posterior distribution of the weight.

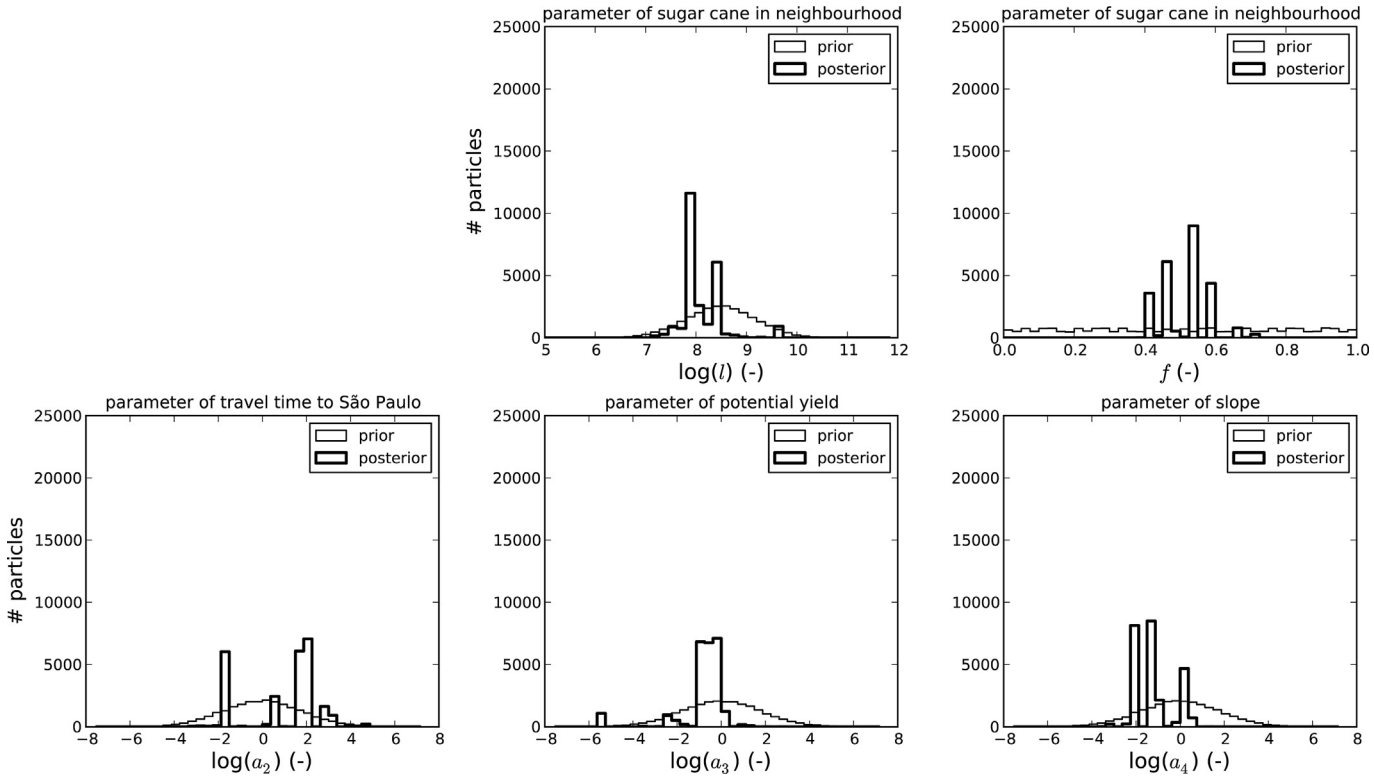

Fig. A2. Prior (thin line) and posterior (thick line) distributions of all parameters for the new block division.

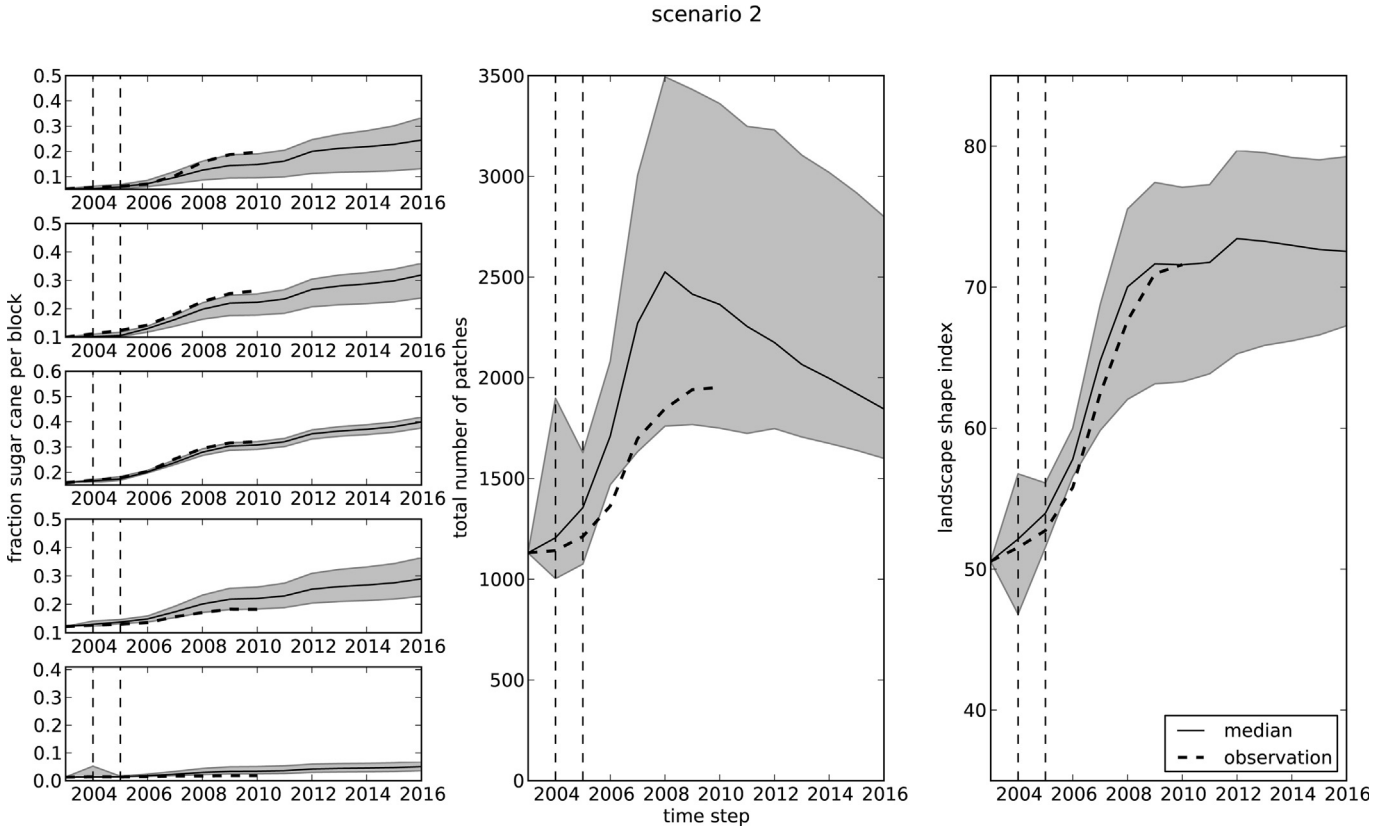

Fig. A3. Comparison of modelled medians (solid lines), 95% confidence intervals (grey areas) and observations (dashed lines) of the spatial metrics for the validation blocks for scenario 2 (particle filter). The filter time steps are indicated with vertical dashed lines.

References

Aerts, J.C.J.H., Goodchild, M.F., Heuvelink, G.B.M., 2003. Accounting for spatial uncertainty in optimization with spatial decision support systems. Trans. GIS 7, 211–230.

Aguiar, A.P.D., Câmara, G., Escada, M.I.S., 2007. Spatial statistical analysis of land-use determinants in the Brazilian Amazonia: exploring intra-regional heterogeneity. Ecol. Model. 209, 169–188.

Aguiar, D.A., Rudorff, B.F.T., Silva, W.F., Adami, M., Mello, M.P., 2011. Remote sensing images in support of environmental protocol: monitoring the sugarcane harvest in São Paulo state, Brazil. Remote Sens. 3, 2682–2703.

Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. 50, 174–188.

Batty, M., 2012. Building a science of cities. Cities 29, S9–S16.

Batty, M., 2005. Agents, cells, and cities: new representational models for simulating multiscale urban dynamics. Environ. Plann. A 37, 1373–1394.

Bengtsson, T., Bickel, P., Li, B., 2008. Curse-of-dimensionality Revisited: Collapse of the Particle Filter in Very Large Scale Systems. In: Probability and Statistics: Essays in Honor of David A. Freedman 2, pp. 316–334.

Berjak, S.G., Hearne, J.W., 2002. An improved cellular automaton model for simulating fire in a spatially heterogeneous Savanna system. Ecol. Model. 148, 133–151.

Bettencourt, L.M.A., 2013. The origins of scaling in cities. Science 340, 1438–1441.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, Oxford, UK.

Collin, A., Bernardin, D., Sero-Guillaume, O., 2011. A physical-based cellular automaton model for forest-fire propagation. Combust. Sci. Technol. 183, 347–369.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.

Csillag, F., Boots, B., 2005. Toward comparing maps as spatial processes. In: Fisher, P. (Ed.), Developments in Spatial Data Handling; 11th International Symposium on Spatial Data Handling. Springer, Berlin Heidelberg, pp. 641–652.

Dai, E., Wu, S., Shi, W., Cheung, C.-., Shaker, A., 2005. Modeling change-pattern-value dynamics on land use: an integrated GIS and artificial neural networks approach. Environ. Manage. 36, 576–591.

de Souza Soler, L., Verburg, P.H., 2010. Combining remote sensing and household level data for regional scale analysis of land cover change in the Brazilian Amazon. Reg. Environ. Change 10, 371–386.

Diogo, V., van der Hilst, F., van Eijck, J., Faaij, A., Verstegen, J.A., Hilbert, J., Carballo, S., Volante, J., 2013. Combining empirical and theory-based land use modelling approaches to assess future availability of land and economic potential for sustainable biofuel production: Argentina as a case study. Renew. Sustain. Energy Rev. (in press).

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10,143–10,162.

Fang, S., Gertner, G., Wang, G., Anderson, A., 2006. The impact of misclassification in land use maps in the prediction of landscape dynamics. Landscape Ecol. 21, 233–242.

Fargione, J., Hill, J., Tilman, D., Polasky, S., Hawthorne, P., 2008. Land clearing and the biofuel carbon debt. Science 319, 1235–1238.

Feng, Y., Liu, Y., Tong, X., Liu, M., Deng, S., 2011. Modeling dynamic urban growth using cellular automata and particle swarm optimization rules. Landscape Urban Plann. 102, 188–196.

Grimm, V., Railsback, S.F., 2012. Pattern-oriented modelling: a 'multi-scope' for predictive systems ecology. Phil. Trans. R. Soc. B Biol. Sci. 367, 298–310.

Hansen, H.S., 2012. Empirically derived neighbourhood rules for urban land-use modelling. Environ. Plann. B Plann. Design 39, 213–228.

Hartig, F., Calabrese, J.M., Reineking, B., Wiegand, T., Huth, A., 2011. Statistical inference for stochastic simulation models – theory and application. Ecol. Lett. 14, 816–827.

Hellmann, F., Verburg, P.H., 2011. Spatially explicit modelling of biofuel crops in Europe. Biomass Bioenergy 35, 2411–2424.

Hellmann, F., Verburg, P.H., 2010. Impact assessment of the European biofuel directive on land use and biodiversity. J. Environ. Manage. 91, 1389–1396.

Hiemstra, P.H., Karssenberg, D., van Dijk, A., de Jong, S.M., 2012. Using the particle filter for nuclear decision support. Environ. Modell. Softw. 37, 78–89.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688.

ICONE, 2012. ICONE – Institute for International Trade Negotiations, Brief Description for the Brazilian Land Use Model – BLUM.

Ivanovic, R.F., Freer, J.E., 2009. Science versus politics: truth and uncertainty in predictive modelling. Hydrol. Process. 23, 2549–2554.

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled Seamless SRTM Data V4.

Jazwinksi, A.H., 1970. Stochastic Processes and Filtering Theory. Academic Press, New York.

Jeremiah, E., Sisson, S.A., Sharma, A., Marshall, L., 2012. Efficient hydrological model parameter optimization with Sequential Monte Carlo sampling. Environ. Modell. Softw. 38, 283–295.

Johnson, B.R., 2010. Eliminating the mystery from the concept of emergence. Biol. Phil. 25, 843–849.

Karssenberg, D., Schmitz, O., Salamon, P., de Jong, K., Bierkens, M.F.P., 2010. A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. Environ. Modell. Softw. 25, 489–502.

Kéfi, S., Rietkerk, M., Alados, C.L., Pueyo, Y., Papanastasis, V.P., ElAich, A., De Ruiter, P.C., 2007. Spatial vegetation patterns and imminent desertification in Mediterranean arid ecosystems. Nature 449, 213–217.

Kocabas, V., Dragićević, S., 2007. Enhancing a GIS cellular automata model of land use change: bayesian networks, influence diagrams and causality. Trans. GIS 11, 681–702.

Lapola, D.M., Schaldach, R., Alcamo, J., Bondeau, A., Koch, J., Koelking, C., Priess, J.A., 2010. Indirect land-use changes can overcome carbon savings from biofuels in Brazil. Proc. Natl. Acad. Sci. U. S. A. 107, 3388–3393.

Lauf, S., Haase, D., Hostert, P., Lakes, T., Kleinschmit, B., 2012. Uncovering land-use dynamics driven by human decision-making – a combined model approach using cellular automata and system dynamics. Environ. Modell. Softw. 27–28, 71–82.

Li, X., Yeh, A.G., 2002. Neural-network-based cellular automata for simulating multiple land use changes using GIS. Int. J. Geogr. Inf. Sci. 16, 323–343.

Liu, X., Li, X., Liu, L., He, J., Ai, B., 2008. A bottom-up approach to discover transition rules of cellular automata using ant intelligence. Int. J. Geogr. Inf. Sci. 22, 1247–1269.

Macedo, I.C., Seabra, J.E.A., 2008. Mitigation of GHG emissions using sugarcane bioethanol (Chapter 4). In: Zuurbier, P., van de Vooren, J. (Eds.), Sugarcane Ethanol: Contributions to Climate Change Mitigation and the Environment. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 95–110.

Manson, S.M., 2007. Challenges in evaluating models of geographic complexity. Environ. Plann. B Plann. Design 34, 245–260.

Nassar, A.M., Rudorff, B.F.T., Antoniazzi, L.B., Aguiar, D.A., Bacchi, M.R.P., Adami, M., 2008. Prospects of the sugarcane expansion in Brazil: impacts on direct and indirect land use changes (Chapter 3). In: Zuurbier, P., van de Vooren, J. (Eds.), Sugarcane Ethanol: Contributions to Climate Change Mitigation and the Environment. Wageningen Academic Publishers, Wageningen, pp. 63–112.

Padua Junior, A.L., Costa Pasini, A.C., Comatsu, C.E., Casarin, D.C.P., Michelino, G.G., von Glhen, H.C., da Silva, I.X., de Moreas, J.F.L., de Carvalho, J.P., Sandoval, M., Valeriano, M., Araujo, N., Brunini, O., Vedovelo, R., Viegas, R., Campaign, R.C.F., Adami, S.F., 2012. Agro-enviromental Zoning – Green Ethanol – Environmental System for São Paulo. Goverment of São Paulo.

Page, S.E., 2011. Diversity and Complexity. Princeton University Press, Princeton, N.J.

Parker, D.C., Hessl, A., Davis, S.C., 2008. Complexity, land-use modeling, and the human dimension: fundamental challenges for mapping unknown outcome spaces. Geoforum 39, 789–804.

Pasetto, D., Camporese, M., Putti, M., 2012. Ensemble Kalman filter versus particle filter for a physically-based coupled surface-subsurface model. Adv. Water Resour. 47, 1–13.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.

Pebesma, E.J., Wesseling, C.G., 1998. Gstat: a program for geostatistical modelling, prediction and simulation. Comput. Geosci. 24, 17–31.

Pijanowski, B.C., Alexandridis, K.T., Müller, D., 2006. Modelling urbanization patterns in two diverse regions of the world. J. Land Use Sci. 1, 83–108.

Pijanowski, B.C., Brown, D.G., Shellito, B.A., Manik, G.A., 2002. Using neural networks and GIS to forecast land use changes: a Land Transformation Model. Comput. Environ. Urban Syst. 26, 553–575.

Rasmussen, R., Hamilton, G., 2012. An approximate bayesian computation approach for estimating parameters of complex environmental processes in a cellular automata. Environ. Modell. Softw. 29, 1–10.

Rudorff, B.F.T., Aguiar, D.A., Silva, W.F., Sugawara, L.M., Adami, M., Moreira, M.A., 2010. Studies on the rapid expansion of sugarcane for ethanol production in São Paulo State (Brazil) Using Landsat Data. Remote Sens. 2, 1057–1076.

Rykiel Jr., E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Model. 90, 229–244.

Salamon, P., Feyen, L., 2009. Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. J. Hydrol. 376, 428–442.

Santé, I., García, A.M., Miranda, D., Crecente, R., 2010. Cellular automata models for the simulation of real-world urban processes: a review and analysis. Landscape Urban Plann. 96, 108–122.

Searchinger, T.D., Heimlich, R., Houghton, R.A., Dong, F., Elobeid, A., Fabiosa, J., Tokgoz, S., Hayes, D., Yu, T., 2008. Use of U.S. croplands for biofuels increases greenhouse gases through emissions from land-use change. Science 319, 1238–1240.

Sparovek, G., Barretto, A.G.d.O.P., Berndes, G., Martins, S., Maule, R., 2009. Environmental, land-use and economic implications of Brazilian sugarcane expansion 1996–2006. Mitigation Adapt. Strat. Global Change 14, 285–298.

Sparovek, G., Berndes, G., Barretto, A.G.d.O.P., Klug, I.L.F., 2012. The revision of the Brazilian Forest Act: increased deforestation or a historic step towards balancing agricultural development and nature conservation? Environ. Sci. Policy 16, 65–72.

Sparovek, G., Berndes, G., Egeskog, A., de Freitas, F.L.M., Gustafsson, S., Hansson, J., 2007. Sugarcane ethanol production in Brazil: an expansion model sensitive to socio-economic and environmental concerns. Biofuels Bioprod. Biorefin. 1, 270–282.

Spiller, E.T., Budhiraja, A., Ide, K., Jones, C.K.R.T., 2008. Modified particle filter methods for assimilating Lagrangian data into a point-vortex model. Physica D 237, 1498–1506.

Straatman, B., White, R., Engelen, G., 2004. Towards an automatic calibration procedure for constrained cellular automata. Comput. Environ. Urban Syst. 28, 149–170.

Torquato, S.A., 2006. Cana-de-açúcar para indústria: o quanto vai precisar crescer. In: Análises e Indicatores do Agronegócio 1.

Tóth, G., Kozlowski, B., Prieler, S., Wiberg, D., 2012. Global Agro-ecological Zones (GAEZ v3.0).
van der Hilst, F., Verstegen, J.A., Karssenberg, D., Faaij, A.P.C., 2012. Spatio-temporal land use modelling to assess land availability for energy crops – illustrated for Mozambique. Global Change Biol. Bioenergy 4, 859–874.
van der Kwast, J., Canters, F., Karssenberg, D., Engelen, G., Van De Voorde, T., Uljee, I., De Jong, K., 2011. Remote sensing data assimilation in modeling urban dynamics: objectives and methodology. In: Procedia Environmental Sciences 7: Spatial Statistics 2011: Mapping Global Change, pp. 140–145.
van Leeuwen, P.J., 2009. Particle filtering in geophysical systems. Mon. Weather Rev. 137, 4089–4114.
van Leeuwen, P.J., 2003. A variance-minimizing filter for large-scale applications. Mon. Weather Rev. 131, 2071–2084.
Verburg, P.H., De Koning, G.H.J., 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, 45–61.
Verburg, P.H., Schot, P.P., Dijst, M.J., Veldkamp, A., 2004. Land use change modelling: current practice and research priorities. GeoJournal 61, 309–324.
Verburg, P.H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., Mastura, S.S.A., 2002. Modeling the spatial dynamics of regional land use: the CLUE-S model. Environ. Manage. 30, 391–405.
Verstegen, J.A., Karssenberg, D., van der Hilst, F., Faaij, A.P.C., 2012. Spatio-temporal uncertainty in spatial decision support systems: a case study of changing land availability for bioenergy crops in Mozambique. Comput. Environ. Urban Syst. 36, 30–42.
von Braun, J., 2008. Rising food prices: what should be done? EuroChoices 7, 30–35.
Walter, A., Dolzan, P., Quilodrán, O., De Oliveira, J.G., Da Silva, C., Piacente, F., Segerstedt, A., 2011. Sustainability assessment of bio-ethanol production in Brazil considering land use change, GHG emissions and socio-economic aspects. Energy Policy 39, 5703–5716.
Yu, J., Chen, Y., Wu, J., Khan, S., 2011. Cellular automata-based spatial multi-criteria land suitability simulation for irrigated agriculture. Int. J. Geogr. Inf. Sci. 25, 131–148.
Zhang, Y., Li, X., Liu, X., Qiao, J., 2011. The CA model based on data assimilation. Yaogan Xuebao J. Remote Sens. 15, 475–482.