Submitted to the Annals of Applied Statistics

SUPPLEMENTARY MATERIAL FOR "MARKOV RANDOM FIELD MODELS FOR VECTOR-BASED REPRESENTATIONS OF LANDSCAPES "

BY PATRIZIA ZAMBERLETTI, JULIEN PAPAÏX, EDITH GABRIEL, THOMAS OPITZ

1. Gibbs sampler.

1.1. *General setting.* We implement a Markov Chain Monte Carlo algorithm of Gibbs sampler type (Gibbs–MCMC) to iteratively simulate a discrete Markov chain whose stationary distribution corresponds to the target model (e.g., Casella and George, 1992), where the configuration of the allocated land-use categories $\boldsymbol{x}$ is the state variable of the system.

The main steps of Gibbs–MCMC are as follows:

i) we define an initial state $\boldsymbol{x}^{(0)}$; and then iteratively run through steps ii and iii:
ii) we generate a new state $\tilde{\boldsymbol{x}}$ given the current state $\boldsymbol{x}^{(j)}$, selecting a component of the vector $\tilde{\boldsymbol{x}^{(j)}}$ and sampling its category update from the distribution of that component conditioned on all the other components sampled so far.
iv) finally, after I_0 iterations, return the configuration $\boldsymbol{x}^{(I_0)}$.

If we need more than one realization of the landscape, we can either run several chains in parallel, or we may run a single chain but return a sample given by the states indexed by $I_0 + \ell I$, $\ell = 1, 2, \ldots$, with the burn-in period $I_0 > 0$ and $L - 1$ intermediate states left out to avoid autocorrelation in the final sample. Since the parameter vector $\boldsymbol{\beta}$ of the model is fixed during each MCMC run, the intractable normalizing constant $c(\boldsymbol{\beta})$ in the probability mass function of the model cancels out in the conditional probabilities used for Gibbs sampling. During the iterations we have to update the calculation of the set of landscape descriptors for each new configuration $\tilde{\boldsymbol{x}}$. With respect to the choice of the initial state $\boldsymbol{x}^{(0)}$ of the system, we have to ensure that $p(\boldsymbol{x}^{(0)}) > 0$, and that valid Gibbs sampling paths $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(j)}$ to and from any configuration $\boldsymbol{x}^{(j)}$ with $p(\boldsymbol{x}^{(j)}) > 0$ are proposed with strictly positive probability. All of the models presented in the paper satisfy $p(\boldsymbol{x}) > 0$ for all possible configurations $\boldsymbol{x}$, such that any initial state is valid. In more general cases, hereditary properties when moving between configurations must be checked. We initialize the system state either at random by drawing the category of an object o_i among all ℓ_i available categories, or by attributing a single category to each object type, or by using an observed configuration from real data.

1.2. *Detailed description of the algorithm.* Here, we present in detail the steps of the algorithm for the iterative but simultaneous simulation of patches and linear elements: at each iteration i, we randomly select an element type (i.e. patch, layer a, or linear element, layer b) to be updated, and then we select an object of this type, followed by a category selection conditioned to all the other elements. If there is more than one time step, i.e., if there are temporal dynamics with time steps $\tau = 1, ..., H_\tau$ and time horizon $H_\tau \in \mathbb{N}$, we also choose at random one of the time steps to be updated.

As outlined above, we propose to use an algorithm where at successive steps for the network, we select at random one of the object of this type and the new category of the element are sampled from the distribution of that component conditioned on all other components sampled so far. The new category is sampled among the available categories conditioned to the allocated categories over the other elements. We denote the full configuration of categories during iteration step i of the algorithm by $\boldsymbol{x}^{(i)}$.

Given

- I, the total number of iterations, and ,
- time steps $\tau = 1, ..., H_\tau$, with H_τ the time horizon, where H_τ is fixed to 1 in the case of purely spatial simulation,

the algorithm for landscape configuration works as follows:

1. Initialize patches in network C as $\boldsymbol{x}_\tau^{C,(0)} = \boldsymbol{x}^{C,(0)}$ with an initial configuration $\boldsymbol{x}^{C,(0)}$, and initialize linear elements in network H as $\boldsymbol{x}_\tau^{H,(0)} = \boldsymbol{x}^{H,(0)}$ with an initial configuration $\boldsymbol{x}^{H,(0)}$, for $\tau = 1, \ldots, H_\tau$.
2. set $i = 1$;
3. while $i \leq I$,
   - select at random the element type type among C and H: sample $U \sim \mathrm{Unif}(0, 1)$; if $U < 0.5$ then $\mathrm{type} = C$ else $\mathrm{type} = H$;
   - if $H_\tau > 1$ select a random the time step: $\tau \sim \mathrm{Unif}(\{1, \ldots, H_\tau\})$;
   - select at random one of the n^{type} object of network type: $J \sim \mathrm{Unif}(\{1, \ldots, n^{\mathrm{type}}\})$;
   - sample the new category for the selected element:

$$\tilde{x}_{J,\tau}^{\mathrm{type},(i)} \sim p\left(\tilde{x}_{J,\tau}^{\mathrm{type},(i)} | \boldsymbol{x}_{(-J)}^{\mathrm{type},(i)}\right)$$

     and denote the full configuration with the new category as $\tilde{\boldsymbol{x}}$;
   - increment $i \leftarrow i + 1$;
4. return the final configuration $\boldsymbol{x}^{(I)}$.

1.3. *Choice of burn-in period.* The landscape descriptors T_k are sufficient statistics in our models of *exponential family type* (Grelaud et al., 2009), i.e., they contain all the information on $\boldsymbol{\beta}$ that we can draw from an observation $\boldsymbol{x}$. Therefore, we can monitor the convergence of Markov chains to their stationary distribution by checking the m simulated series $T_k^{(j)}$, $k = 1, \ldots, m$, through trace plots, which further allows us to determine an appropriate burn-in period, and to analyze the mixing behavior of the Markov chains to fix the number $I - 1$ of intermediate states to be left out (see, e.g. Kiêu et al., 2013). In practice, we have found that the number of iterations needed for burn-in depends on the combination of size of the landscape and complexity of the model, and especially on the type of landscape descriptors involved. The running time necessary to simulate one landscape in one Markov chain for the examples discussed in this paper ranges from several seconds to several minutes.

For illustration, we here report trajectories of the landscape descriptors in a rather complex model for 50000 MCMC iterations with the small domain D1 in Figure 2, and for 1 million MCMC iterations with the large domain D3 in Figure 3, used to check the algorithm convergence.

2. Temporal landscape descriptor. In this section, we define the temporal landscape descriptor and give an example of temporal dynamics with configurations correlated over consecutive time steps, i.e., we evaluate temporal interactions. An example specification is:

$$T_{temp}^C(\boldsymbol{x}) = \sum_{i=1}^{n^C} \sum_{\tau=2}^{H_\tau} t(x_{i,\tau}^C, x_{i,\tau-1}^C). \tag{1}$$

Here, T_{temp}^C captures time dynamics for network C of crops over a horizon of H_τ discrete time steps.

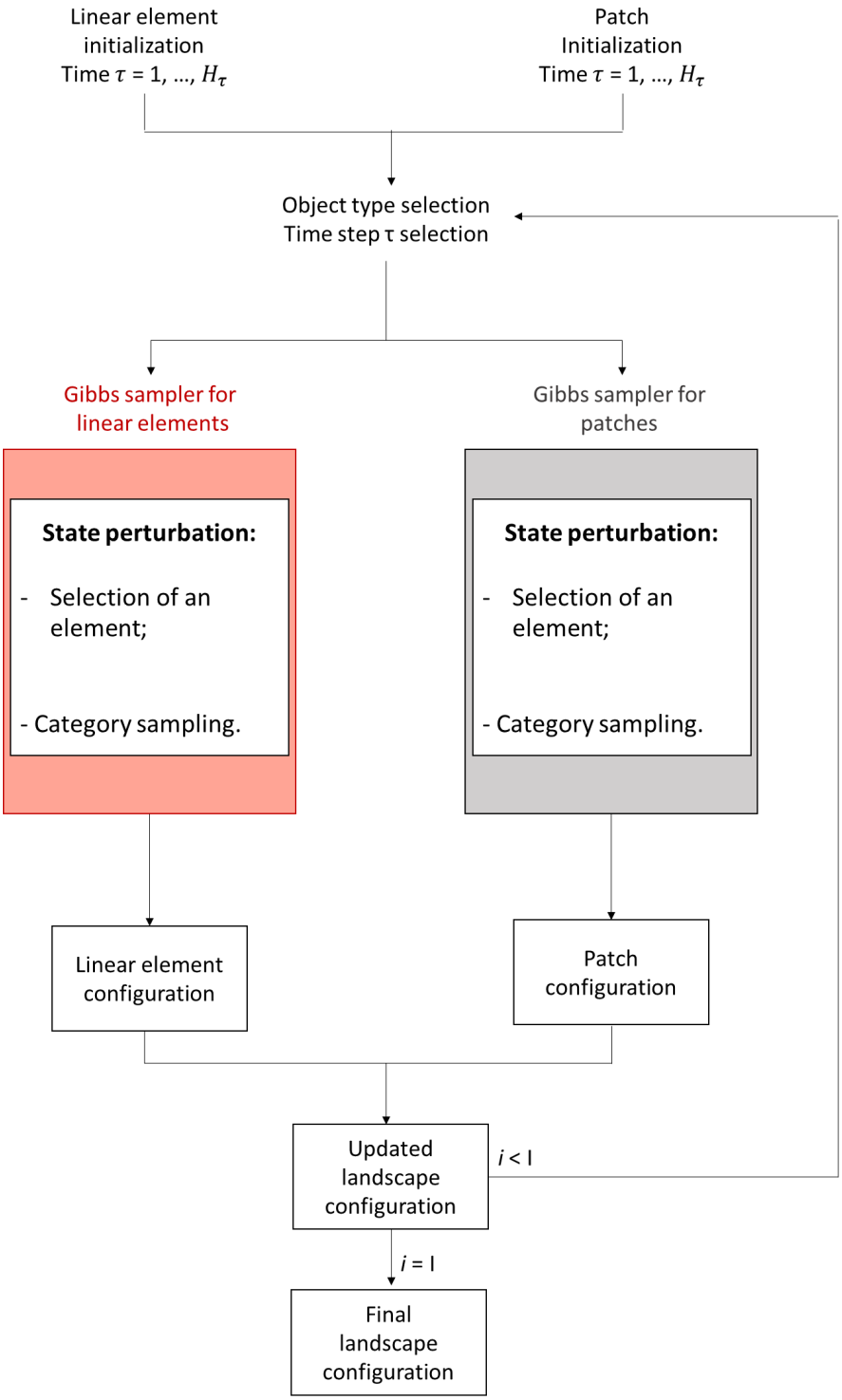

FIG 1. *Schematic representation of the Gibbs sampler. The index i indicatess the current iteration, I is the fixed number of iterations, τ is the current time, and H_τ is the time horizon.*

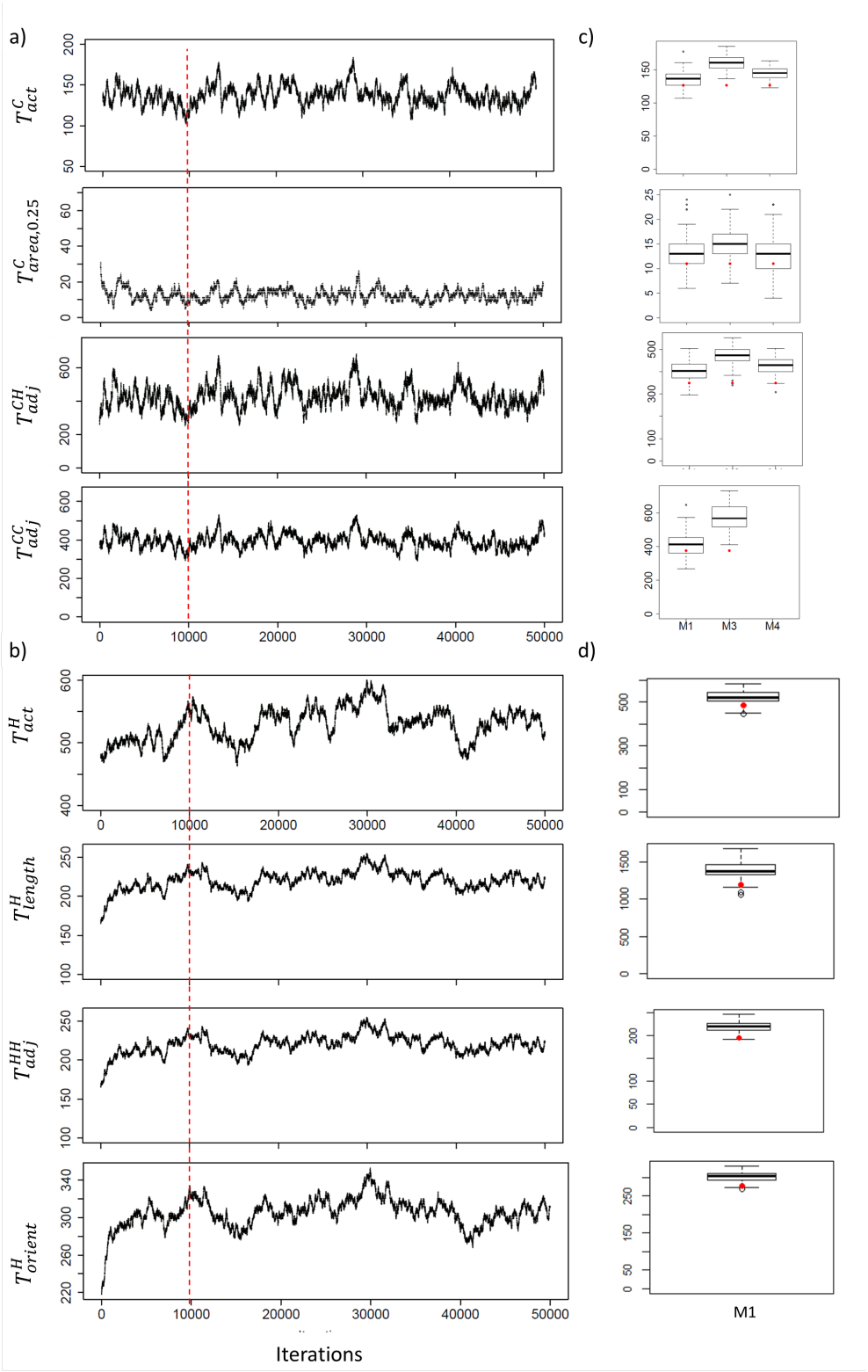

FIG 2. *MCMC simulation of a model for domain for D1. Convergence diagnostics in Panels a,b: trace plot examples for patch allocation with crop (a), and for linear segment allocation with hedges (b); the red dashed lines show the selected burn-in. Panels c, d: landscape descriptor values in domain D1 for 100 simulated landscapes (boxplots) and the real landscape (red dot) using models M1, M3 and M4 for crop and M1 for hedge; for patches (c) and linear elements (d). Descriptors in panels a,c from top to bottom: small crop area ($T^{C}_{\mathrm{area},0.25}$), crop-hedge adjacency (T^{CH}_{adj}), crop-crop adjacency (T^{CC}_{adj}). Descriptors in panels b,d from top to bottom: long hedge allocation (T^{H}_{length}), hedge-hedge adjacency (T^{HH}_{adj}), allocation of horizontally oriented hedges (T^{H}_{orient}).*

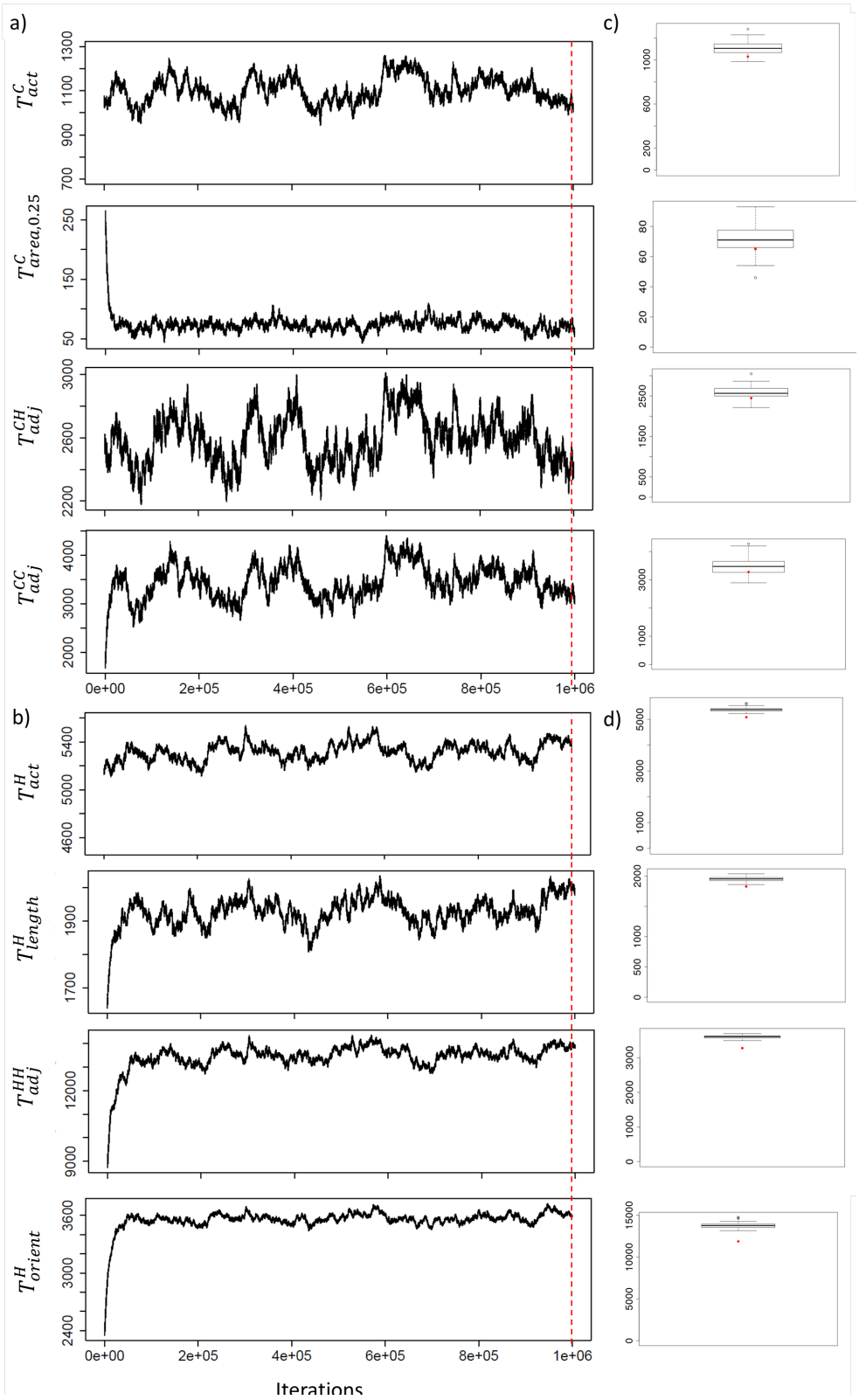

FIG 3. *Convergence diagnostic for D3 with model M1: trace plot for patch allocation with crop (a), linear segment allocation with hedges (b). The landscape descriptor values for the real landscape (red dot) and for the simulated ones (box plot) at the end of the iterations for patches (c) and linear elements (d). Panel a) and c) from top to bottom: small crop area ($T^{C}_{\text{area},0.25}$), crop-hedge adjacency (T^{CH}_{adj}), crop-crop adjacency (T^{CC}_{adj}). Descriptors in panels b,d from top to bottom: long hedge allocation (T^{H}_{length}), hedge-hedge adjacency (T^{HH}_{adj}), allocation of horizontally oriented hedges (T^{H}_{orient}).*

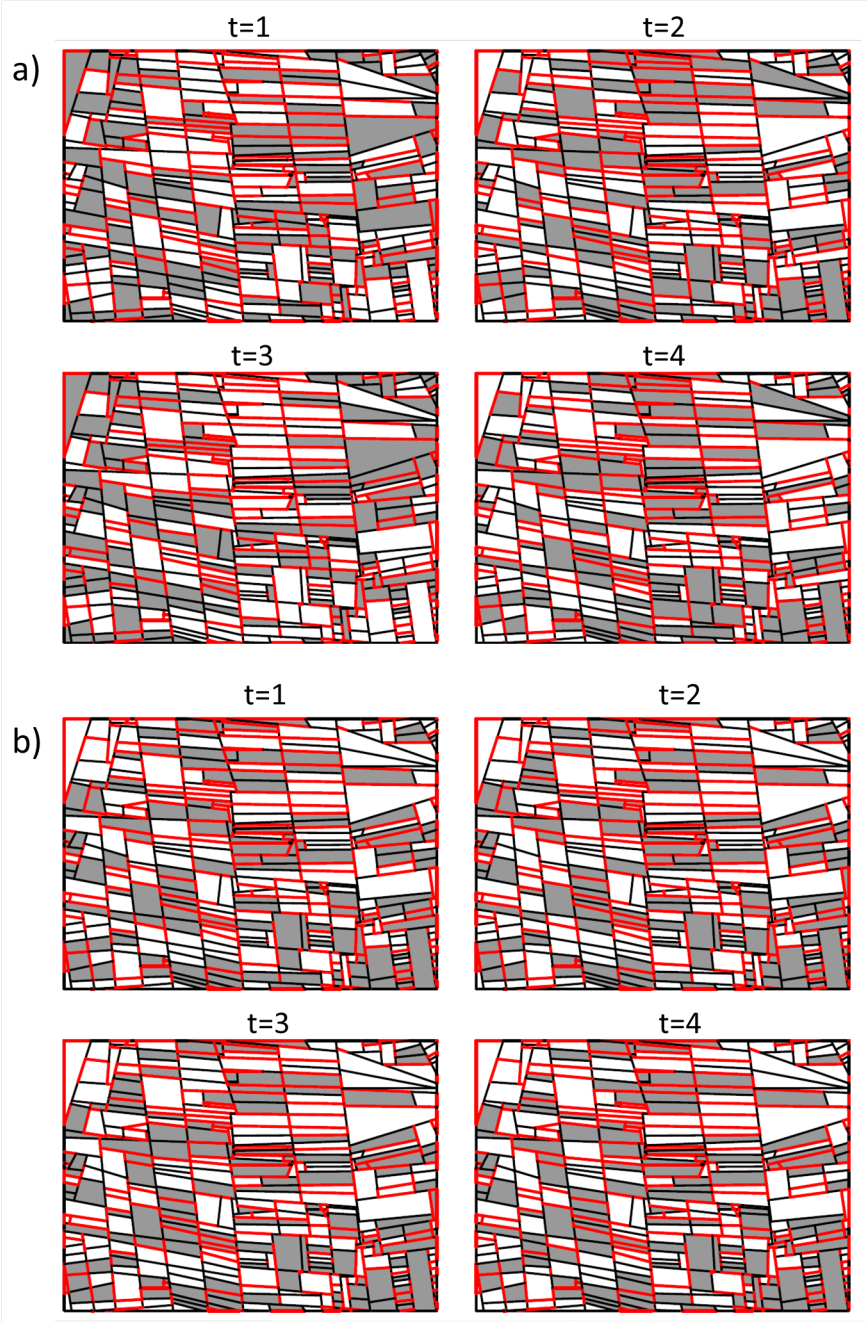

FIG 4. *Examples of crop rotation simulations with positive time correlation (Panel a) and negative time correlation (Panel b). Simulations are performed over four years. Gray boxes stand for crop ($x_i^C = 1$), white boxes stand for semi-natural habitat ($x_i^C = 0$). Red lines stand for hedges that are not influenced by rotation and are kept fixed.*

The specification of the temporal interaction among two objects of C aims to simulate crop rotation using 2 allocation categories (i.e., $x_i \in \{0, 1\}$) : *crop* ($x_i^C = 1$) or *natural habitat* ($x_i^C = 0$). Its formulation is given by:

$$t(x_{i,\tau}^C, x_{i,\tau-1}^C) = 1(x_{i,\tau-1}^C = x_{i,\tau}^C)$$

Simulation examples are presented in Figure 4.

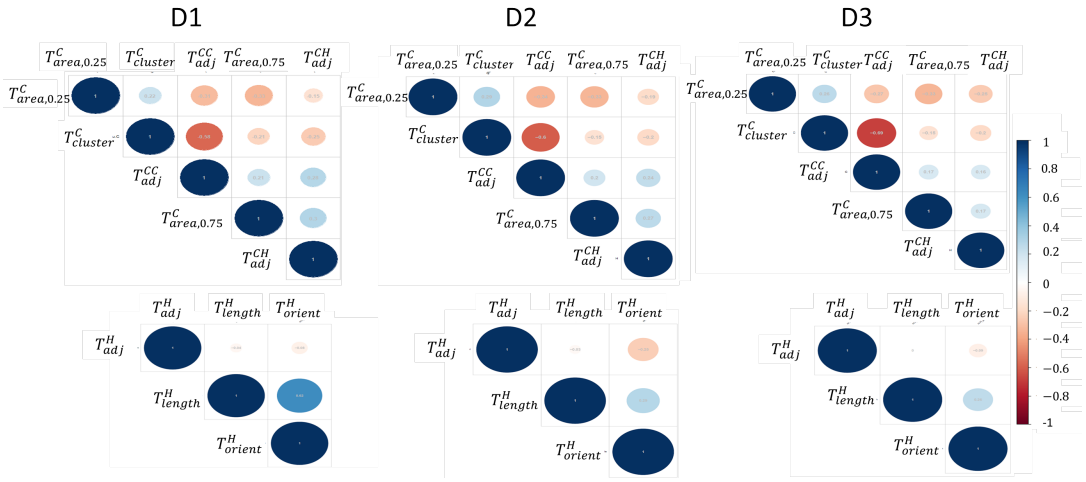

FIG 5. *Correlation of landscape descriptors for the spatial domains D1, D2, D3 for the crop network C (upper panel) and the hedge network H (lower panel).*

3. Landscape descriptor correlation. In order to avoid including strongly correlated landscape descriptors in a landscape model, we check the correlation among covariates arising in the logistic regression. In Figure 5 we display the correlation among all landscape descriptors for the three spatial domains.

4. Full results for variograms. To assess model diagnostics, we compute empirical variograms by transforming our landscapes in rasters. We contrast the variogram of the real landscape with the simulated ones. In Figure 6, variograms for each of the 3 spatial domains with model M1 are shown. They are discussed in the main text where we put focus on model comparison over the spatial domain D1.

5. Complete results for M1, M3 and M4 in the spatial domain D1. Here, we present all results of modelss M1, M3, M4 for the spatial domain D1. In Figure 7a, the boxplot of parameter estimate are shown for crop and hedge categories. In Figure 8, there is a focus on the crop category comparing estimated parameters for M1, M3 and M4. Specific values are reported and discussed in the main text in Table 4. For models M3 and M4, validation results related to network metrics are shown in Figure 9, while results for raster metrics are shown in Figure 10 and Figure 11 for M3 and M4, respectively. Lastly, in Figure 12 we show the variation of the residual standard deviation of model M1 to highlight the improvement achieved by the introduction of the large area landscape descriptor; more details are also provided in the main text.

6. Complete results for model M1 and all domains. In Figure 7 we report the boxplots for parameter estimates based on 100 simulations of the fitted model M1 for the three study domains.

6.1. *Model M1-D2.* The complete validation results for the network metrics in the study area D2 are reported in Figure 13. Figure 14 shows landscape metrics. Figure 17b illustrates the inter-connection metrics.

In Table 1, we report results for the Monte–Carlo pseudo p-values for network scale metrics and raster-based landscape metrics.

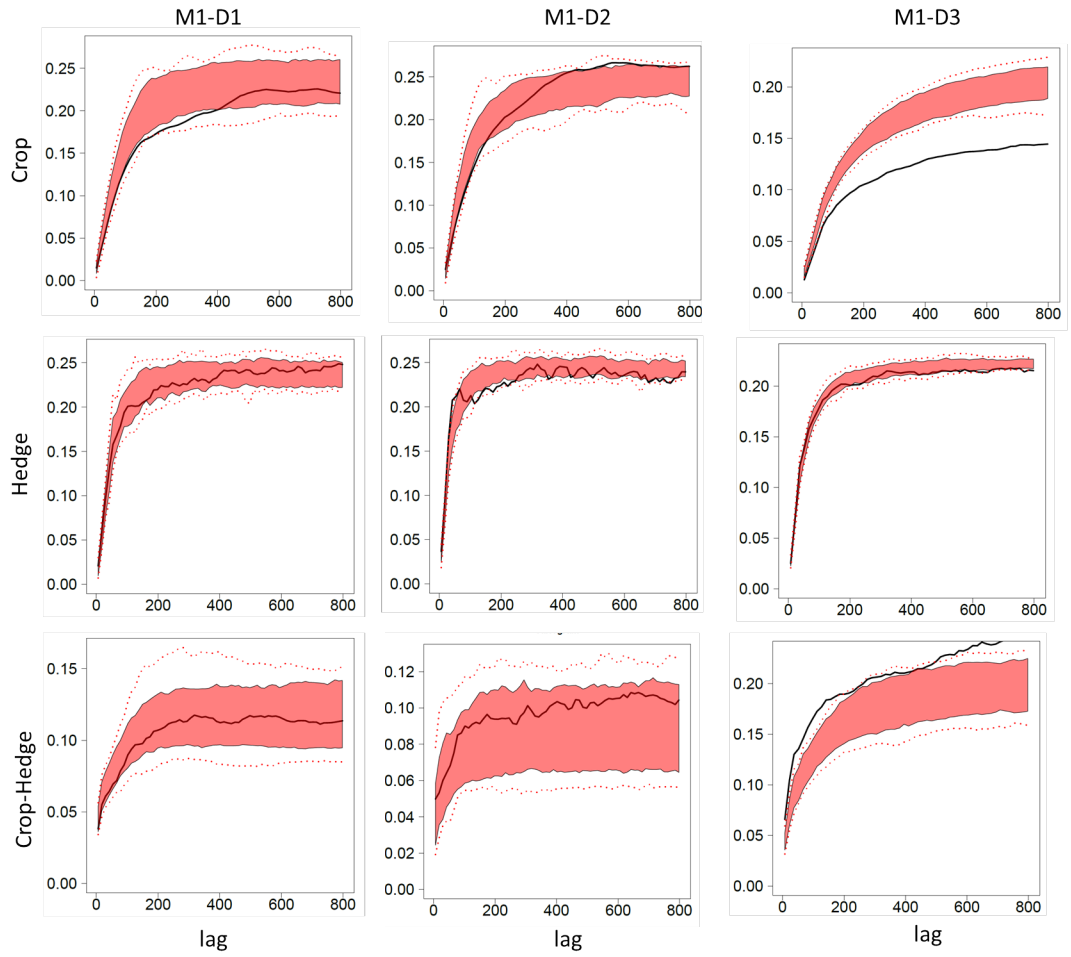

FIG 6. *Variograms for the* 3 *spatial domains D1, D2, D3.*

TABLE 1
Pseudo p-values obtained through the Monte–Carlo statistical test for the domain D2.

	Semi-natural	**Crop**	**Hedge**
Diameter	-	0.04	0.24
Efficiency	-	0.23	0.13
Cluster average	-	0.07	0
PLAND	0.09	0.10	0.11
PD	0	0.19	0.03
PARA_MN	0.04	0	0.12
ENN_MN	0.34	0	0.33
IJI	0.12	0.13	0.01
CLUMPY	0.13	0.21	0.03

6.2. *The model M1-D3.* The complete validation results for the network metrics in the study area D3 are reported in Figure 15. Figure 16 shows landscape metrics. Figure 17c illustrates the inter-connection metrics. The mean of the *Betweenness* distribution in the real landscape and in the simulations lies in the upper region of the boxplot, which is due to the presence of a small number of large outliers not shown in the boxplot.

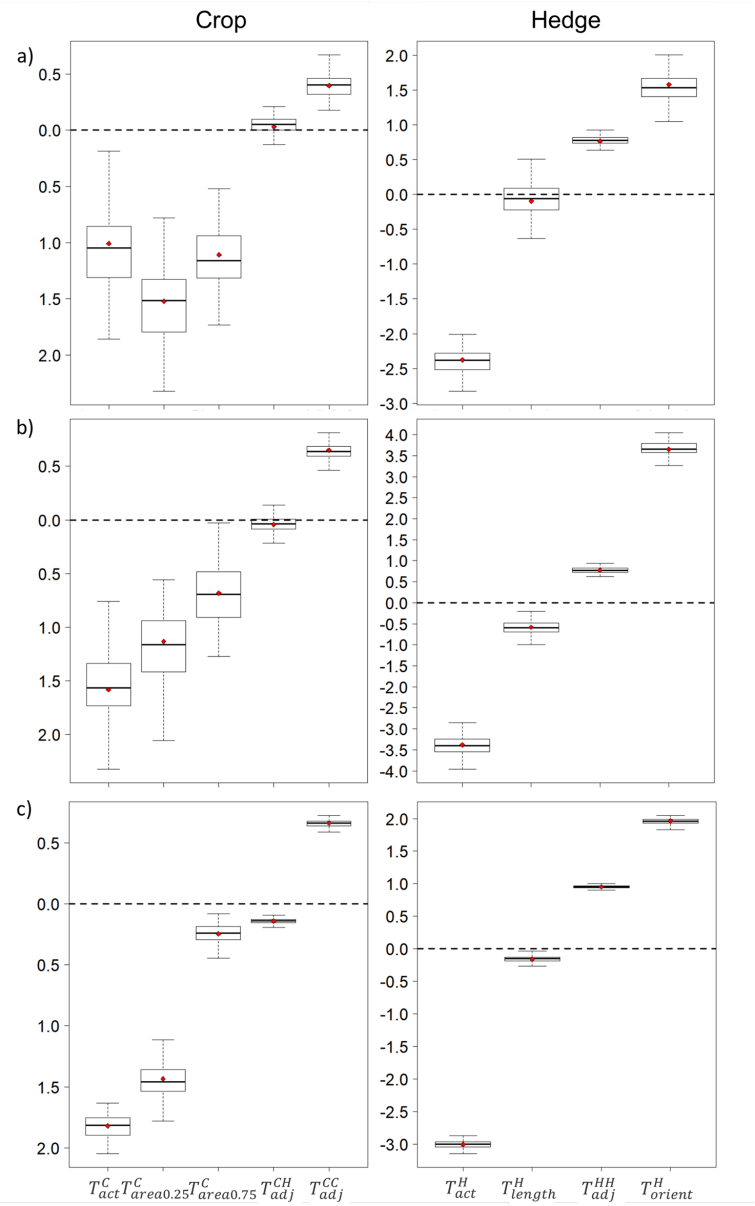

FIG 7. *Parameter estimation for study area D1 (Panel a), D2 (Panel b), D3 (Panel b) with model M1. Left: crop network; right: hedge network. Red dots represent the estimated value. Boxplots represent 100 simulations.*

In Table 2 we report results for the Monte–Carlo pseudo p-values for network-scale metrics and raster-based landscape metrics.

7. Pseudo-p-values for network metrics. The numbers in Table 3 report the proportion $p \in [0, 0.5]$ of the simulated metric values that are "more excentric" than the observed one; e.g., if the observed value is below the median and 26 among the 100 simulated values are even lower, we report 0.26. These *pseudo-p-values* imply that observed metrics for the crop network still appear realistic under the model. Overall, network-scale results indicate slightly stronger clustering of crop in the model as compared to reality, but still with similar order of magnitude for metric values. We also report pseudo-p-values for hedge network-scale metrics in Table 3, which show stronger discrepancy between observed and simulated values. Node-

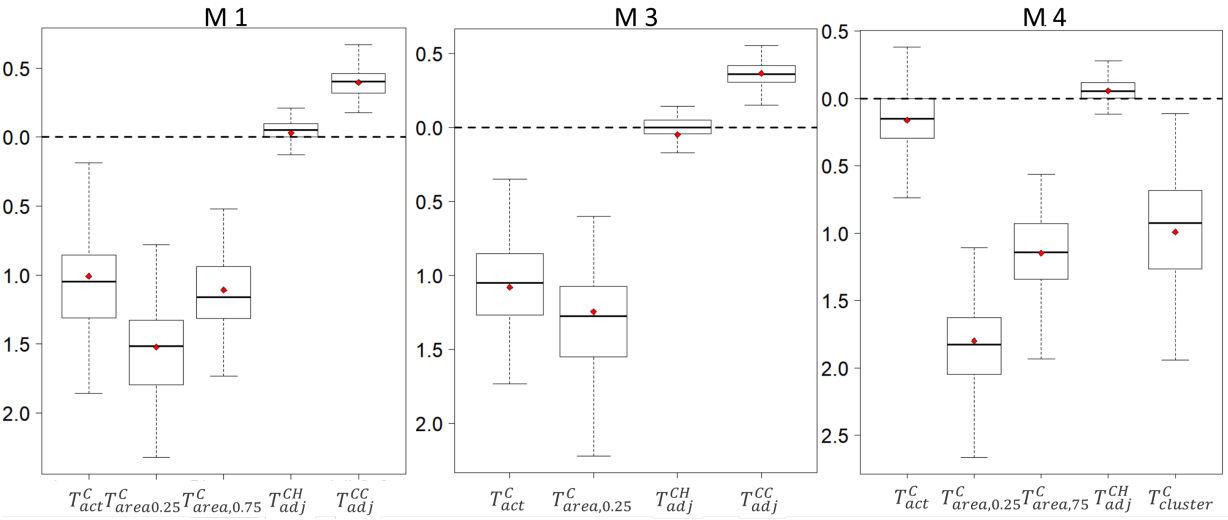

FIG 8. *Parameter estimation for the crop network in domain D1 with models M1, M3 and M4 (by columns). Red dots represent the estimated value. Boxplots represent 100 simulations.*

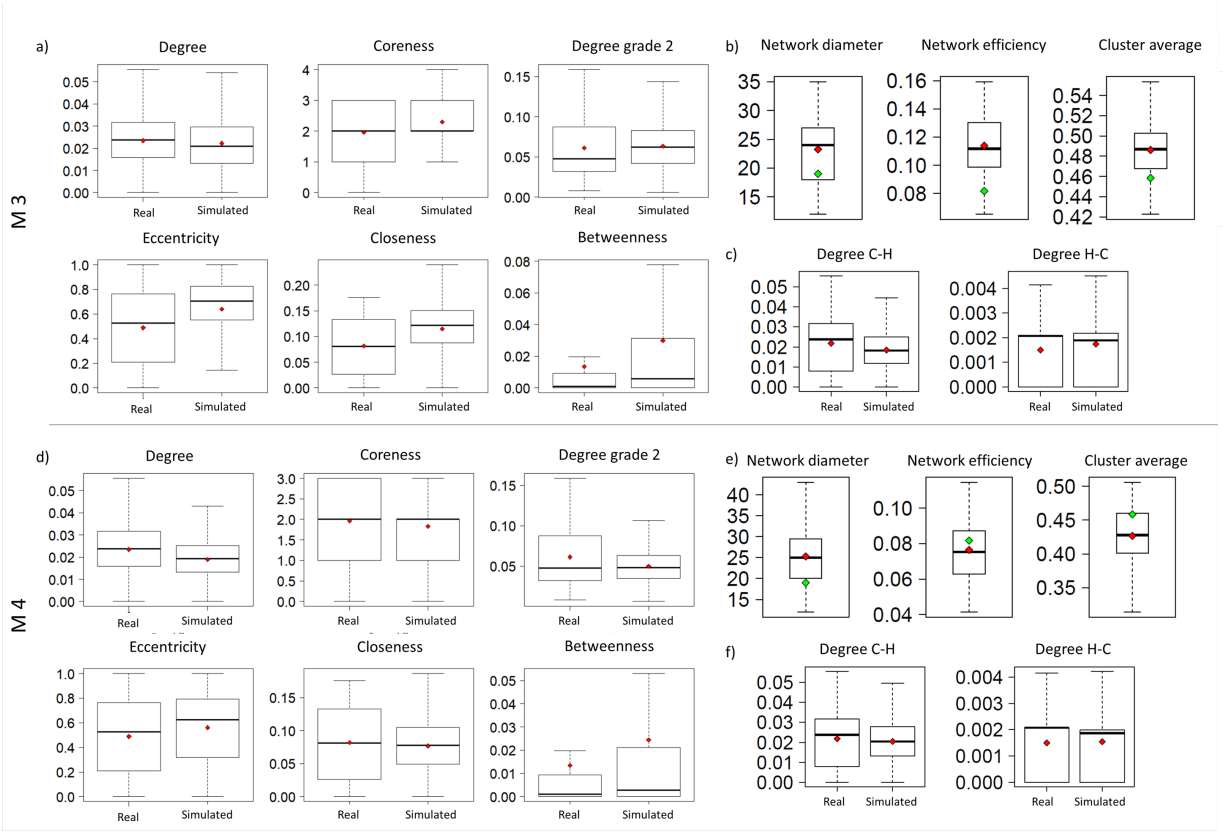

FIG 9. *Validation of crop network metrics in D1 with models M3 and M4 at node scale (panels a,d), at network scale (panels b,e) and for the degree among crop and hedges (panels c,f). In panels a,c,d,f, boxplots represent the distribution of the node metrics for the real landscape network (left boxplots) and the for the simulated landscapes (right boxplots). Red dots represent mean values of the node metric distribution of the real and simulated networks, resspectively. In panels b,e, boxplots represent distributions of the simulated landscapes; red dots represent mean values of the simulated landscapes; green dots represent the real landscape network values.*

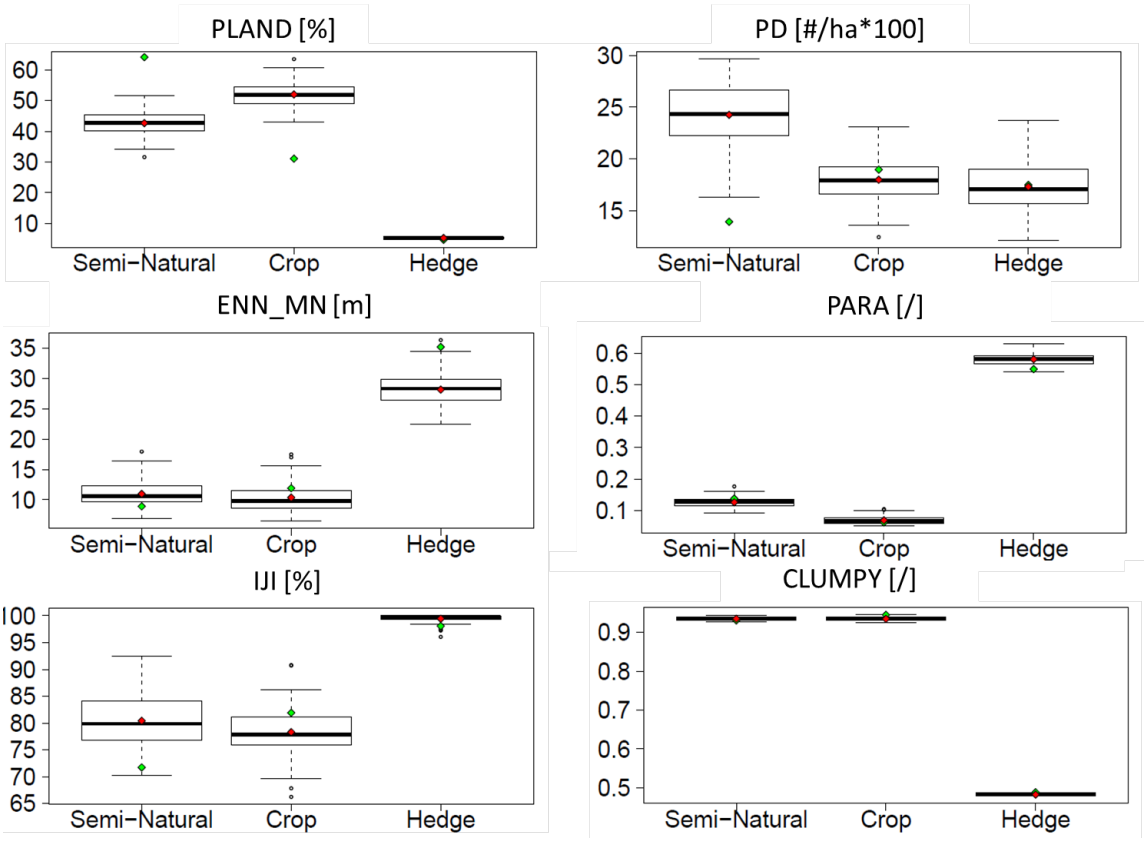

FIG 10. *Validation results for raster metrics in the domain D1 for model M3. Boxplots represent simulated landscapes transformed to raster format for the three habitats (i.e., semi-natural, crop, hedges). Red dots represent mean values of each habitat for simulated landscapes. Green dots represent values of each habitat for the real landscape.*

TABLE 2
Pseudo p-values obtained through the Monte–Carlo statistical test for the large domain D3.

	Semi-natural	**Crop**	**Hedge**
Diameter	-	0.30	0.15
Efficiency	-	0.35	0.15
Cluster average	-	0.34	0
PLAND	0	0	0
PD	0	0.01	0.41
PARA_MN	0	0	0
ENN_MN	0	0.01	0.16
IJI	0	0	0.05
CLUMPY	0.36	0	0

scale metrics for hedges, more directly controlled through the network descriptors included in our model, remain satisfying.

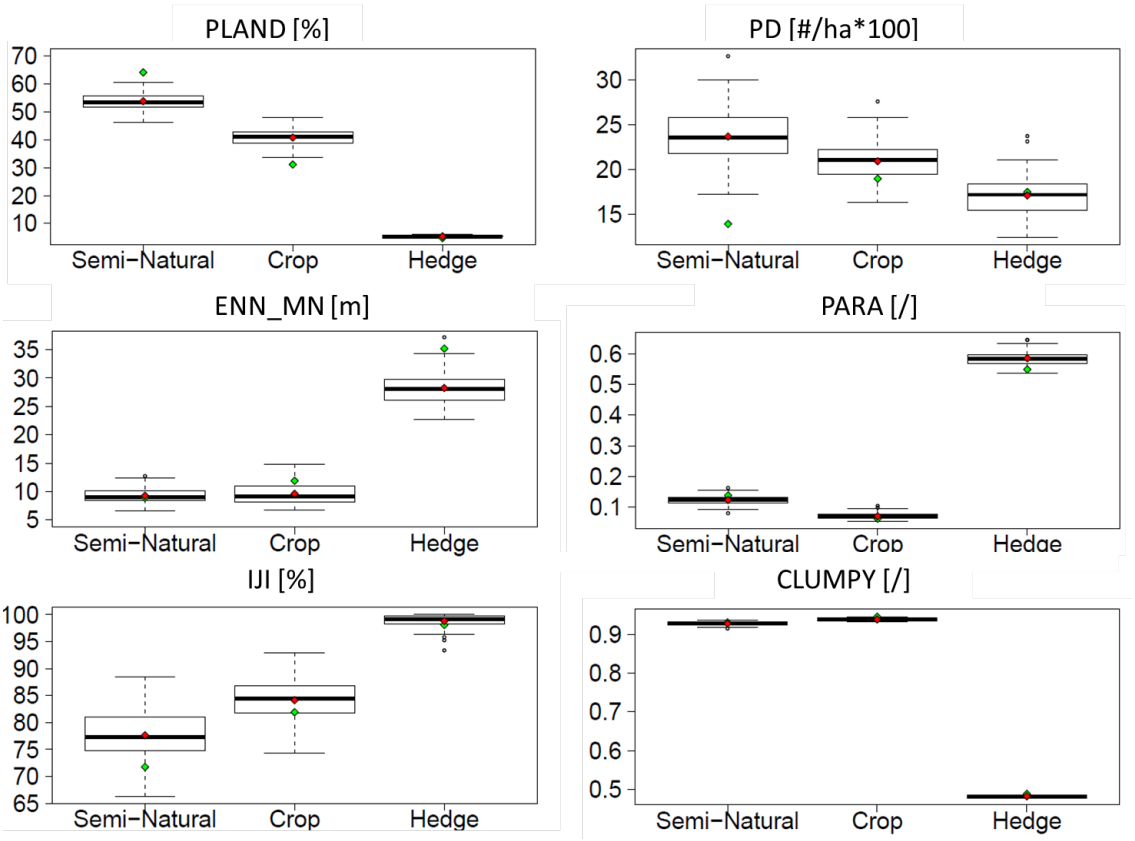

FIG 11. *Validation results for raster metrics in the domain D1 for model M4. Boxplots represent simulated landscapes transformed to raster format for the three habitat types (i.e., semi-natural, crop, hedges). Red dots represent mean values of each habitat for simulated landscapes. Green dots represent values of each habitat for the real landscape.*

TABLE 3
Pseudo-p-values of network-scale metrics and raster-based metrics for D1 and crop models M1, M3, M4.

	Semi-natural			**Crop**			**Hedge**
	M1	M3	M4	M1	M3	M4	M1
Diameter	-	-	-	0.57	0.26	0.15	0.19
Efficiency	-	-	-	0.56	0.06	0.38	0.23
Cluster average	-	-	-	0.16	0.13	0.28	0
PLAND	0.08	0	0	0.10	0	0	0
PD	0	0	0	0.44	0.26	0.19	0.37
PARA_MN	0.20	0.20	0.17	0.12	0.33	0.24	0.06
ENN_MN	0.49	0.11	0.43	0.30	0.24	0.14	0.01
IJI	0.45	0.02	0.09	0.47	0.19	0.28	0.47
CLUMPY	0.19	0.11	0.26	0.24	0.02	0.02	0

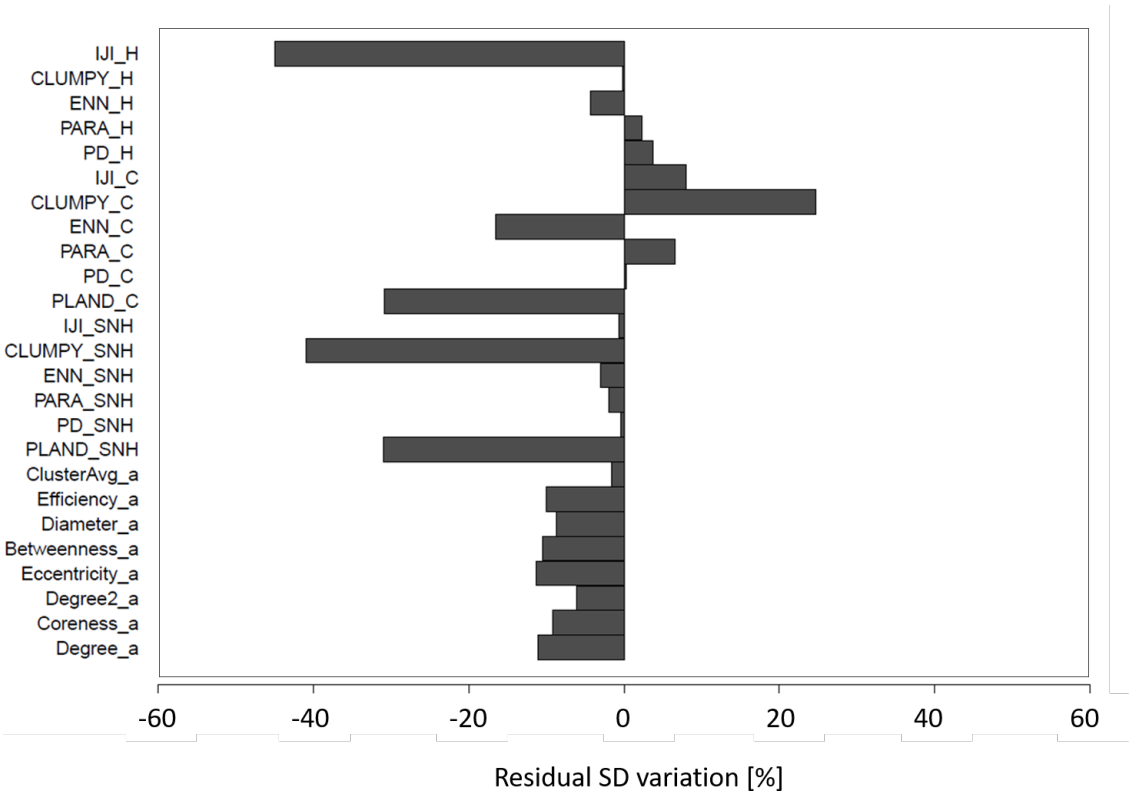

FIG 12. *Percentage variation of the residual standard deviation (SD) of model M1 in domain D1 with respect to model M3in domain D1. The letter a refers to the network patch network. Regarding raster metrics, SNH stands for Semi-natural habitat, C stands for Crop and H stands for hedges.*

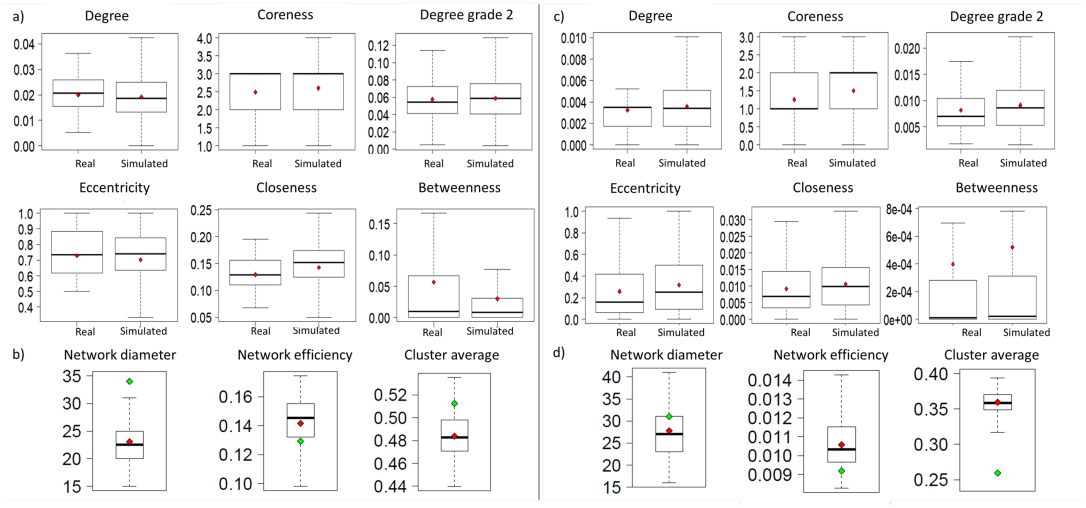

FIG 13. *Validation of network metrics for the domain D2. Validation at node scale (panels a,c), at network scale (panels b,d) for crop network (left) and hedge network (right), respectively. In panels a,c, boxplots represent distributions of node metrics for the real landscape network and for simulated landscapes. Red dots represent mean values of the node metric distribution of the real and simulated networks. In panels b,d, boxplots represent simulated landscapes. Red dots represent mean values of the simulated landscapes. Green dots represent the real landscape.*

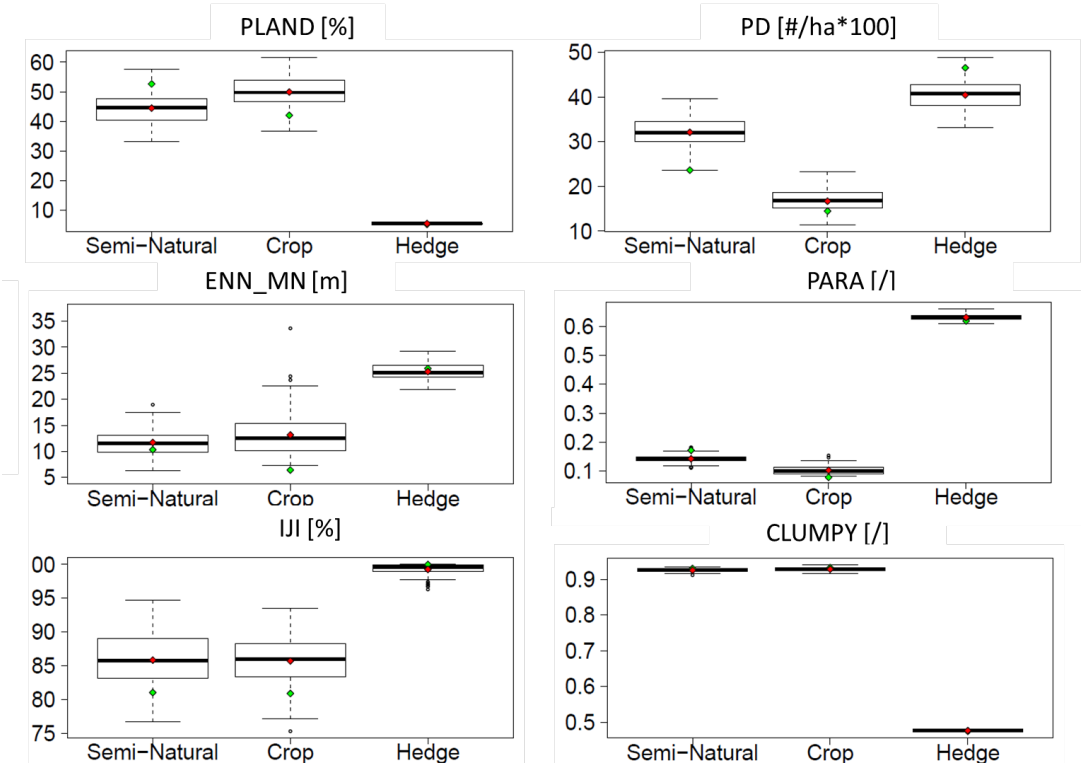

FIG 14. *Validation of raster metrics for the domain D2. Boxplots represent the simulated landscapes in raster format for the three habitat types (i.e., semi-natural, crop, hedge). Red dots represent mean values of each habitat for simulated landscapes. Green dots represent values of each habitat for the real landscape.*

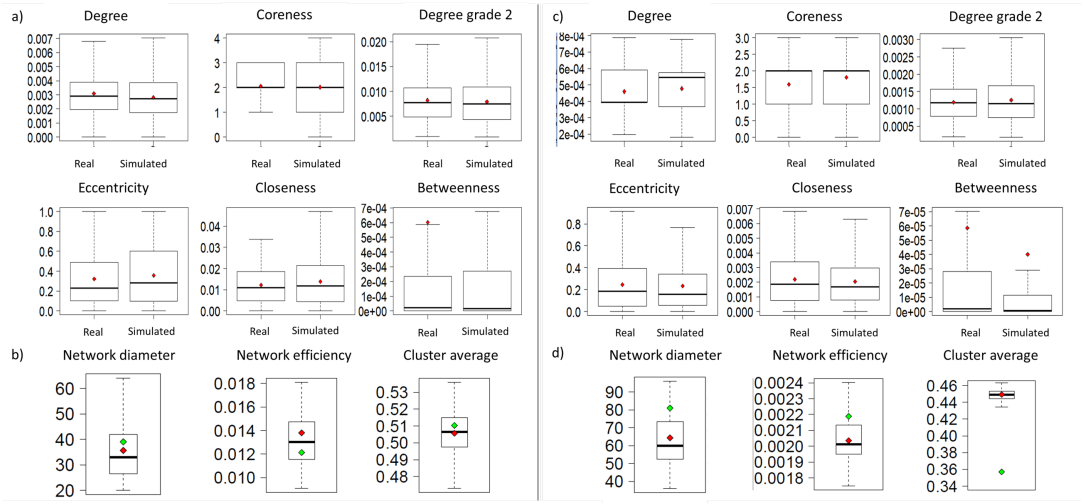

FIG 15. *Validation of network metrics for the domain D3. The description is the same as in Figure 13.*

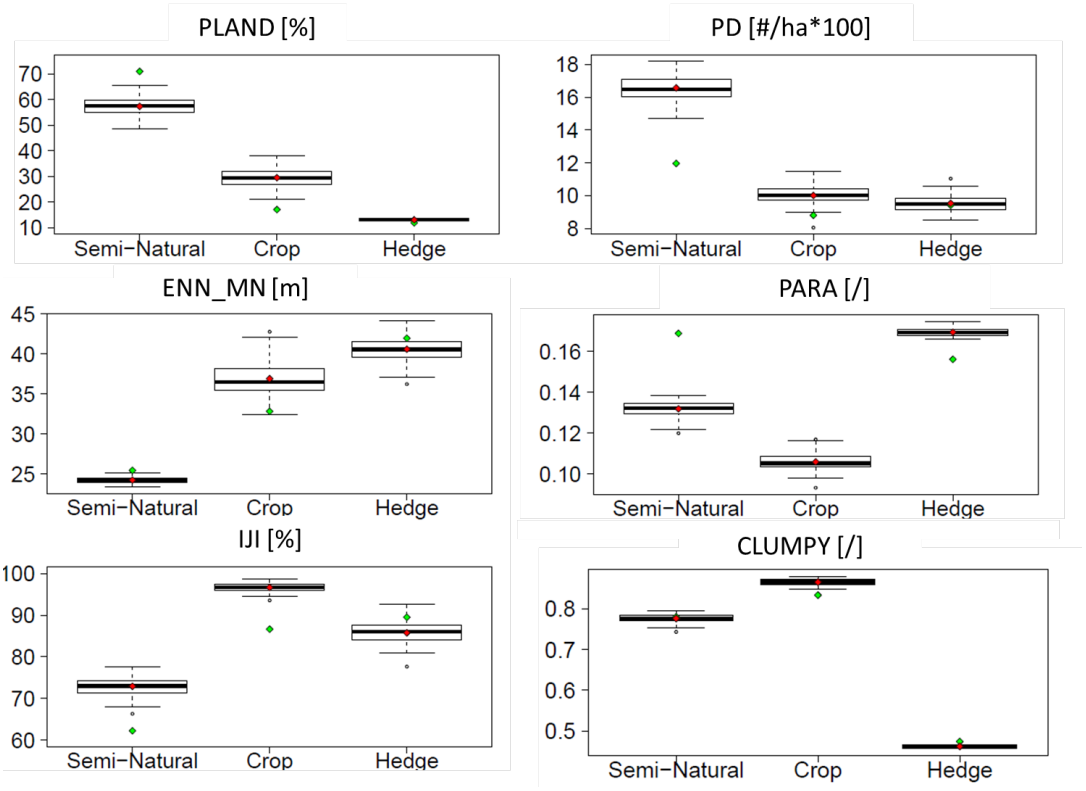

FIG 16. *Validation of landscape metrics for the domain D3. The description is the same as in Figure 14.*

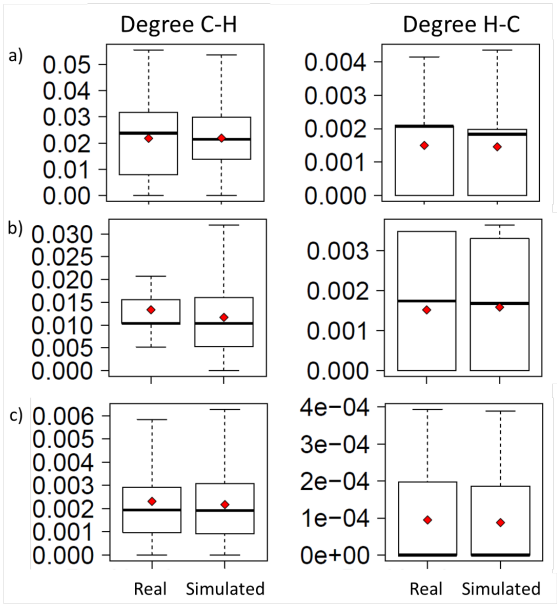

FIG 17. *Validation of network metrics for the three domains D1, D2 and D3 (panels a,b,c, respectively) related to inter-connections in the multi-layer network with model M1. In each display, boxplots show the distribution of the node metric for the real landscape (left boxplot) for the simulated landscapes (right boxplot) for the Crop-to-Hedge degree, which counts the number of links from crop patches to hedges (left column), and the Hedge-to-Crop degree, counting the number of links from hedges to crop patches (right column).*

REFERENCES

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F., Taly, J.-F., et al. (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4(2):317–335.

Kiêu, K., Adamczyk-Chauvat, K., Monod, H., and Stoica, R. S. (2013). A completely random T-tessellation model and Gibbsian extensions. *Spatial Statistics*, 6:118–138.