Supplementary file S7 to: Grimm, V. et al. (2020) 'The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism' *Journal of Artificial Societies and Social Simulation* 23 (2) 7: http://jasss.soc.surrey.ac.uk/23/2/7.html [10.18564/jasss.4259]

S7: Describing simulation experiments

[1]Grimm et al. (2010) pointed out that in most cases it is necessary to include in "Materials and Methods", following the ODD, a section called "Model Analysis" or "Simulation Experiments". Here we recommend merging the elements related to model analysis and simulation experiments, and adding, where relevant, calibration, into a single section entitled "Calibration, simulation experiments, and model analysis". While this section would not be part of the ODD protocol itself, we provide some basic guidelines to standardize its description.

We suggest several subsections, which correspond to key elements of the model "evaludation" framework and TRACE documents suggested by Augusiak et al. (2014) and Grimm et al. (2014), respectively. Ideally, a TRACE document is produced first, but so far this has only been done for more complex models addressing specific applied questions in ecology (see example TRACE documents in Supplement S6). Still, most elements of TRACE are relevant for any model type and purpose, and the same holds for the categories suggested below. Even if a model only explores an idea or illustrates a narrative, providing a thorough model analysis, including in particular sensitivity and robustness analysis, will make a model more useful.

The main subsections can, depending on the journal's formatting requirements, have their own section titles or be referred to by highlighting these titles in the text via italics or capitalization. In Table 1, we suggest further subsections, which are meant as reminder of relevant issues. Below, for each of these categories and issues, we also compiled example texts from the literature.

1. Calibration and model output verification

This element first describes which parameters were calibrated, why they were, and how they were; for the last of these it is necessary to describe: (i) whether parameters were calibrated in the sub- or full model, and independently or simultaneously, (ii) the range of values tested for each parameter and method used to sample the entire parameter space (e.g., full factorial design, Latin hypercube sampling, optimization methods, Bayesian methods; see Thiele et al. 2014), (iii) initial conditions and simulation settings (e.g., simulation length, spatial landscape, time series of environmental drivers, values of non-calibrated parameters), (iv) empirical patterns to be matched by the model, (v) fitting criteria, i.e. metrics used to quantify how well the model output matches the data (e.g., sum of squared standardized errors) and strategy (e.g., best-fit, categorical calibration), and (vi) other technical details, such as number of replicates of each parameter set, software used to implement the parameter space sampling algorithm or to analyse model fitting, and so on. This section contains a lot of technical information that enables replication, but that might not be critical for understanding what has been done. Details of the technical information might therefore go into supplementary material, which should be provided not only on the journal's website, but also in an open

[1] Lead author of this supplement: Daniel Ayllón

repository together with the ODD. Of course, since this is part of the "Methods" section of a publication, results of calibration are not presented in this section.

Second, this element should ideally also describe the methods or formal tests used to assess model accuracy. Specifically, the element should describe how model outputs, derived from simulations run with the best parameterization, match each pattern used for calibration or model development, together with the quantitative criteria used to decide whether a certain pattern was matched by the model.

2. Model output corroboration

This element provides information on the methods employed to compare real model predictions, i.e. without further calibration, with independent data and patterns that were not used while the model was developed, parameterized, and verified. The kind of information provided is similar to that of the previous element: (i) the empirical observations or theoretical patterns reproduced by the model, (ii) the simulation settings (e.g., simulation length, spatial configuration, the time series of environmental drivers, initial conditions, parameter set) and number of replicates performed, and (iii) the formal tests and criteria used to establish if a certain observation or pattern was accurately reproduced by the model.

We consider the distinction of model output corroboration, which does not include calibration, from model output verification, essential. Unfortunately so far this distinction usually was not made. Often, any model output is called "prediction", which is misleading if the model output was tuned, or calibrated, to match certain patterns. Admittedly, so far few ecological or environmental models make real, or independent prediction for model corroboration, but perhaps this is because modellers often do not try hard enough to explore whether the model predicts certain phenomena or relationship which were not purposefully wired into the model.

3. Sensitivity analysis

This element describes how the influence of varying model inputs on simulation outputs was explored. Sensitivity analyses are most commonly focused on varying model parameters, but initial conditions, input data, model configuration (e.g., spatial arrangement), or submodels are other model components for which sensitivity analysis can be performed. Therefore, it must be explicitly indicated which model components are analysed for sensitivity, and which contrasting conditions are tested on which model outputs.

Regarding sensitivity analysis of model parameters, it is necessary to fully describe the experimental design, indicating (i) whether a local or global analysis was performed, (ii) which parameters were assessed, what parameter values/range was tested or what parameter space sampling method was used in the case of global analyses, (iii) the analysis technique employed (e.g., screening, regression-based, or variance decomposition methods; see Thiele et al. 2014), describing all the details, (iv) settings of the other model components when this information is relevant, and (v) other technical details, such as number of replicates, and software used to implement the statistical algorithms.

4. Simulation Experiments for Model Analysis

Simulation models serve as virtual laboratories. One develops and uses them to answer specific research questions; and for this, one designs simulation experiments (scenarios), very much as one designs experiments in real laboratories (Lorscheid et al. 2012). Accordingly, one should also describe simulation experiments in the same way as one describes real experiments, by stating the purpose of each experiment and by providing all details required to replicate them, including, if not fixed between experiments, the spatial and temporal settings, the list of model inputs that are varied (e.g., parameters, initial values of state variables, time series of environmental drivers, spatial configuration), the number of replicates, the variables observed, and the statistical analysis performed on such observations.

As for the design of the simulation experiments, like in experiments on real systems, one usually keeps most parameters and settings constant and varies only one or a few, in a factorial design (Lorscheid et al. 2012, Thiele et al. 2014). Then, one can turn certain mechanisms on and off, or use extreme settings and stress tests to better understand how the output of the full model emerges. An important class of experiments, which so far are under-represented in the literature, is aimed at trying to 'break' a model; that is, if a model reproduces a certain observation and thereby offers a possible explanation, what are the limits of this explanation? Often we learn more by exploring where a model fails than by tweaking parameters and submodels to make the model output look realistic (Thiele and Grimm 2015). Grimm and Berger (2016) summarized approaches for breaking models, and how we can learn from it, under the term "**Robustness analysis**". Considering its potential to investigate models behaviour and to infer essential processes from it, we suggest to consider it as an explicit part of any model analysis.

Checklist and structure for the documentation of "Calibration, simulation experiments, and model analysis"[2]

1 Calibration and model output verification

- Indicate parameters that were calibrated and the reason why they were so.
- Indicate how they were calibrated, providing information about:
  - Whether parameters were calibrated in the sub- or full model, and independently or simultaneously.
  - The range of values tested for each parameter and method used to sample the parameter space, if applicable.
  - Initial conditions and simulation settings.
  - Empirical patterns to be matched by the model.
  - Fitting criteria and strategy for choosing the optimal parameter set.
  - Other technical details, such as number of replicates of each parameter set, software used to implement the parameter space sampling algorithm or to analyse model fitting.

[2] Section titles are linked to examples. Remember: „ALT + ←/ALT + →" brings you back and forth in the hyperlink chain.

2 Model output corroboration

- Indicate the empirical observations or theoretical patterns to be reproduced by the model.
- Describe the simulation settings and relevant technical details.
- Indicate the formal tests and criteria (statistical, quantitative, qualitative) used to assess model performance and whether validation patterns were accurately reproduced by the model.

3 Sensitivity analysis

- Indicate the model component analysed.
- Describe the experimental design, providing information about:
  - Whether a local or global analysis was performed.
  - Which parameters were assessed, the parameter values/range tested, indicating the parameter space sampling method used in the case of global analyses, and the model outputs examined.
  - The analysis technique employed, describing the details.
  - Settings of the other model components when this information is relevant.
  - Other technical details, such as number of replicates, and software used to implement the statistical algorithms.

4 **Simulation experiments**

- Describe the aim of the experiment.
- Indicate the initial conditions and experiment settings, providing information about:
  - Parameter values/ranges tested (with parameter space exploration algorithm if applicable), including external datasets, such as environmental time series or maps.
  - Model structure (only if different from the default model description).
  - Simulation settings, including time step, simulation length, stop conditions, and number of replicates.
- Regime shifts, events, and other scheduled model forcing, if applicable, explaining:
  - When and to what values parameter sets are changed.
  - Whether particular events are scheduled.
  - Whether there are structural changes forced upon the model runtime at predefined times.
- Variables observed and statistical analyses.

Examples from existing publications:

1 Calibration and model output verification:

Indicate parameters that were calibrated and the reason why they were so.
"The parameters most suitable for calibration are those to which model results are highly sensitive and for which there is little basis, other than calibration, for selecting values. There

are only six parameters that are especially suited for calibration in inSTREAM-Gen. Four of them (*habDriftRegenDist*, *habDriftConc*, *habSearchProd*, and *habPreyEnergyDensity*) are involved in the bioenergetics model and are easily calibrated using observed individual growth and survival rates. The other two parameters, *mortFishAqPredMin* and *mortFishTerrPredMin*, define the daily probability of surviving aquatic and terrestrial predation under the most vulnerable conditions." [TRACE doc from Ayllón et al. 2016]

"Most of the model parameters were assigned using values collected in the field, in the laboratory or in published literature (Table 1). The parameters that we had no field data for were parameterised using POM. This included time of the first passage, the distance between a field from which geese were disturbed and the closest roost, which determines whether geese move to the roost after being disturbed or to a field (termed 'roost-disturbance radius', see section 2.4and Table 2), and the memory factor of the geese (α) (Appendix A, equation A5)." [Chudzinska et al. 2016]

Indicate how they were calibrated, providing information about:
(i) Whether parameters were calibrated in the sub- or full model, and independently or simultaneously.
"For each model component, we estimated parameter values from the literature or via calibration. We calibrated the full model only to ensure that survival and growth rates were reasonable, because these rates directly affect habitat selection. Full-model calibration used observed survival and growth from a 75-d period from mid-July to early October. We calibrated mortality of young-of-the-year (age-0) fish using the aquatic predation-risk parameter and mortality of older fish using terrestrial predation risk. Growth rates of yearling (age 1) and older trout were calibrated with the drift-food-availability parameter, after which growth of age-0 trout was calibrated with the benthic food-availability parameter." [Railsback and Harvey 2002]

(ii) The range of values tested for each parameter and method used to sample the parameter space, if applicable.
"These three parameters were set to the following values: time of the first passage: 1–4 h with 0.5 intervals; roost-disturbance radius: 0.1 -1 km with 0.1 km intervals; alpha 0.001; 0.01; 0.03; 0.05; 0.07; 1. We ran 10 simulations for each of the 420 parameter combinations for each FDR." [Chudzinska et al. 2016]

"We subsequently used a Latin hypercube sampling design (Iványi et al. 1979), optimizing the sample with a genetic algorithm, by means of the *lhs* R package v. 0.10 (Carnell 2012) to draw 2000 parameter sets from the entire parameter space defined by the six parameters selected for calibration." [Ayllón et al. 2016]

(iii) Initial conditions and simulation settings.
"The calibration simulations ran from 1 October 1993 through 30 September 2004. The population initialization parameters were derived from data observed in 1993 and global parameters were set to the values described in Section 3 of the TRACE document (Appendix A)." [Ayllón et al. 2016]

(iv) Empirical patterns to be matched by the model.
"InSTREAM-Gen was calibrated within the pattern-oriented framework (Wiegand et al. 2003, Grimm et al. 2005) by using 12 years (1993-2004) of population data from the Belagua River (see sections 3.2, pages 28-29, and 6 of the TRACE document for further details). We calibrated these parameters by attempting to reproduce six time-series patterns: length-at-age of age-1 trout (L1), age-2 trout (L2), and age-3 and older trout (age-3Plus; L3), and abundance of the same age classes (age-1, -2 and -3Plus; A1-3)." [Ayllón et al. 2016]
"Nine parameters were inversely determined (Grimm and Railsback, 2012; Wiegand et al.,2003) via calibration of two movement patterns obtained from one tracked male (ER11; see section 2.3). The patterns used were *moving duration* and *distance from home* and these were measured from both simulated and telemetry tagged seals. Home represents the location where the seal was captured for telemetry data, but in the model it refers to the initial position of the seal. Distance from home was recorded at each time step, and moving duration was measured when the seal reached haul out site after a foraging bout." [Liukkonen et al. 2018]

(v) Fitting criteria and strategy for choosing the optimal parameter set.
"We used the sum of standardized squared errors (SSSE) to evaluate agreement between the observed and predicted patterns. This quantitative measure is computed as:

$$\sum_i \frac{(\mathrm{sim}_i - \mathrm{obs}_i)^2}{\mathrm{obs}_i},$$

where *sim* and *obs* represent the simulated and observed values for each year *i* of the 1993-2004 time series, measured at September 1.

We next followed a Monte Carlo Filtering approach, by which tested patterns were applied as filters to separate good from bad sets of parameter values (Wiegand et al. 2003, Grimm and Railsback 2005). The first patterns used as filters were lengths-at-age. We considered an observed field length-at-age pattern to be accurately reproduced by a model simulation when SSSE was equal to or less than the sum of yearly deviations corresponding to a maximum of 10% of the observed annual value. Parameter sets passing this filter were then filtered by abundance patterns. We only retained parameter sets producing a median SSSE lower than a value equal to a yearly deviation of 30% of the observed value. We selected the parameter set having the overall lowest SSSE values for tested abundance patterns." [Ayllón et al. 2016]

"For each parameter set, the model was run ten times; means of the output variables' medians were calculated and compared to telemetry data to assess model performance. Error from telemetry data was calculated according to:

$$Error = \frac{|Obs - Sim|}{Obs},$$

where *Obs* is the observed median for the two variables calculated from the telemetry data, while *Sim* is the mean of the model runs' medians of the given output variables.

Based on the first round of simulations and following a filtering approach (Wiegand et al. 2004), the parameter sets having an error for the *moving duration* variable below 15% were selected; among the sets passing this first filter, only the parameter sets presenting a total combined error for both patterns below 500% were retained to determine the range for the second round of simulations. In this stage, the parameter set producing the lowest total error from telemetry data was chosen." [Liukkonen et al. 2018]

"We considered we had a final parameter set when two main criteria were met: (1) the deviance measure (see Appendix C) of the 16 patterns could not be improved by small changes of any of the parameters and (2) when all chi-square tests of the comparison between observed and 30 replicates of simulated fish size distributions of each year for 0+ parrs, 1+ parrs, smolts and anadromous individuals were non-significant." [Piou and Prévost 2012]

(vi) Other technical details, such as number of replicates of each parameter set, software used to implement the parameter space sampling algorithm or to analyse model fitting.
"We subsequently used a Latin hypercube sampling design (Iványi et al. 1979), optimizing the sample with a genetic algorithm, by means of the *lhs* R package v. 0.10 (Carnell 2012) to draw 2000 parameter sets from the entire parameter space defined by the six parameters selected for calibration [...] This process was replicated 5 times." [Ayllón et al. 2016]

2 Model output corroboration (validation):

1) Indicate the empirical observations or theoretical patterns to be reproduced by the model.
"After its calibration, the model was tested against the observed time series of population biomass of age-1 trout (B1) and age-2 and older trout (age-2Plus; B2)." [Ayllón et al. 2016]

"To assess the performance of DisPear, we used its final parameterization (Table 1) to conduct 100 replicates of simulations and compare simulated outputs to 46 observed field patterns describing i) dispersers' movement (10 patterns), ii) dispersers' habitat use (30 patterns), and iii) fruits and faeces abundance and spatial distribution and clustering (6 patterns)." [Fedriani et al. 2018]

"In the second phase, additional data from five adult individuals (HE07♀: 3003 relocations, 191 d, 367.15 m/20 min; KJ07♂: 4475 relocations, 218 d, 518.94 m/20 min; OL10♀: 2155, 180 d, 302.27 m/20 min; TO09♂: 3254 relocations, 194 d, 505.44 m/20 min; and VI09♂: 4618 relocations, 199 d, 281.88 m/20 min) were used for model validation and final calibration of one parameter (see section 2.5)." [Liukkonen et al. 2018]

"We validated the inter-annual hydrodynamics of the landscape (flood-events, sub-system 1) by comparing the simulated frequency of flood events to the frequency measured between 1986 and 2006 at the Middle River Elbe gauging station, available from WSV/BfG (2011). For the validation of landscape response, we compared the simulated mean hydroperiod of the temporary ponds with those expected by the hydroperiod gradient. To validate sub-system 2 (spawning site selection), the simulated number of males with status "calling" or "paired" at each pond (mean value of years with wet or intermediate hydrological conditions out of 100

replicates over 50 years of simulation) was compared with the number of males recorded in the field at each pond in 2010 (a wet year) and 2011 (an intermediate year), as described in Dick et al. (2017). For the validation of population dynamics (sub-system 3), our main working hypothesis was that the moor frog population observed in the Elbe floodplain had to be stable under observed close-to-natural habitat conditions. Therefore, we tested whether the population dynamics reached a stable population size at the same level as that observed in the field over a period of 50 years in 100 replicates. For the validation of demographic structures, we used information from Hartung and Glandt (2008), who stated a female-to-offspring ratio of 1:8.8. Further information on each validation procedure is reported in Appendix A, Section 6." [Dick and Ayllón 2017]

2) Describe the simulation settings and relevant technical details.

"To compare the population dynamics generated by the different models with that of natural vole populations, each model was run 1000 times for 35 yr (or until the population went extinct) and the first 5 yr were discarded (to omit the vole population prior to weasel presence). We monitored the vole population size in week 44 (1 November), which corresponds to the period in autumn when most natural populations are monitored (Stenseth 1999, Krebs 2013) and simulated (Turchin and Hanski 1997)." [Radchuk et al. 2016]

3) Indicate the formal tests and criteria (statistical, quantitative, qualitative) used to assess model performance and whether validation patterns were accurately reproduced by the model.
- *Example of the use of statistical criteria:*
"Bayesian estimation of the probability of the difference between observed and predicted growth rates were calculated using the R package BEST.R." [Phang et al. 2016]

- *Example of the use of quantitative criteria:*
"We calculated the deviation (%) of the simulated output (SO) from the observed pattern (OP) as: (SO – OP)/OP * 100." [Fedriani et al. 2018]

- *Example of the use of qualitative criteria:*
"We then analyzed the output to determine whether the patterns were reproduced. These analyses were generally qualitative and graphical because the patterns are generally qualitative and because the model was uncalibrated." [Railsback and Johnson 2011]

3 Sensitivity analysis:

1) Indicate the model component analysed.
- *Example of sensitivity analysis on model parameters*:
"Although we used empirically collected and literature-based values to build the model, we performed a global sensitivity analysis to evaluate how the patterns emerging from the model were affected by variations in the input parameters. The aim was to decompose the model outputs' variance into variances attributable to each input parameter, but also to evaluate the interaction between parameters." [Chudzinska et al. 2016]

- *Example of sensitivity analysis on initial conditions*:
"Sensitivity of predicted trout abundance to the number of trout at the start of a simulation was investigated by varying the initial abundance in nine otherwise identical scenarios." [Railsback et al. 2009]

- *Example of sensitivity analysis on input spatial configuration*
"This section examines the sensitivity of predicted trout population biomass to site-specific habitat input: the size and spatial arrangement of habitat cells, and the input describing hiding cover, velocity shelter, and spawning gravel in cells." [Railsback et al. 2009]

"To study the effects of the number and location of haul out sites, we ran simulations initialising the model with 120, 240, 480, 960 and 1920 rocks. Haul out sites were either kept in the same locations between replicates or randomly distributed at the beginning of each replicate." [Liukkonen et al. 2018]

- *Example of sensitivity analysis on input time series*:
"We conducted a global sensitivity analysis to identify those model parameters with the strongest influence on model outputs under two water temperature scenarios representing current non-stressful and projected climate-change stressful temperatures. We analysed parameter sensitivity under two temperature scenarios because parameters controlling effects of high temperature on reproduction, survival or metabolism may have little effect under current conditions when temperatures are never extreme but could have strong effects at projected higher temperatures." [Ayllón et al. 2016]

- *Example of sensitivity analysis on specific submodels*:
"The model was then simulated on 125 × 128 landscape to examine how a single female territory size varies with respect to habitat quality, i.e., cell-based prey biomass [...] Reproduction and mortality processes were turned off." [Carter et al. 2015]

2) Describe the experimental design, providing information about:
(i) Whether a local or global analysis was performed.
- *Example of the application of global sensitivity analysis*:
"We conducted a global sensitivity analysis to identify the model parameters with the strongest influence on model outputs." [Liukkonen et al. 2018]

- *Example of the application of local sensitivity analysis*
"We conducted local sensitivity analyses to identify the parameters with the strongest influence on Global Fst values." [Baggio et al. 2018]

"Sensitivity analyses were carried out for the default setting when *varroa* mites were added (10 virus-free and 10 virus-carrying mites on day 0 of the simulation). Sixty-one parameters were tested individually, as testing the number of parameter combinations necessary for a full global sensitivity analysis is not possible within a realistic time-scale." [Becher et al. 2014]

(ii) Which parameters were assessed, the parameter values/range tested, indicating the parameter space sampling method used in the case of global analyses, and the model outputs examined.

"[...] we followed a two-step protocol: (1) screening 72 selected model parameters to differentiate influential and non-influential parameters (remaining parameters were cast aside based on a first pre-analysis and results from previous sensitivity analyses described in Railsback et al. 2009; see TRACE document) [...]

All 72 screened parameters were varied over five levels according to predefined ranges, the central value being the value used to calibrate the model (Table A19 in Section 7 of Appendix A) [...]

The sensitivity analysis examined seven model outputs: mean total abundance and biomass of both young-of-the-year (YOY; age-0) and older (age-1 and older) trout, and the mean genotypic values of length at emergence and length maturity threshold (for both males and females) of breeders over a 12-year period." [Ayllón et al. 2016]

"We selected seven parameters that varied over the following ranges: Exh_{adult} (exhaustion-rate-adults; 0.01-0.25), sl_{adult} (mean-speed-adults; 10-600), $sl\mathrm{SD}_{adult}$ (sd-speed-adults; 10-600), ta_{adult} (mean-turning; -10 - +10), $ta\mathrm{SD}_{adult}$ (sd-turning; 10-100), M_R (ref-mem-decay-rate; 0-1), and *Vis* (land-distance; 1-10) [...] The sensitivity analysis examined two model outputs: *moving duration* and *distance from home*." [Liukkonen et al. 2018]

"We conducted local sensitivity analyses to identify the parameters with the strongest influence on Global Fst values. In each local analysis, the selected parameter was varied over levels shown in Table 2, while the rest of parameters were set to their standard values (see section 2.3)." [Baggio et al. 2018]

"Each parameter was multiplied by a factor ranging from 0.1 to 4 (Table 1), except when the default value was 0 or an integer value was required (details of the sensitivity analyses are given Appendix S4, Supporting information). *Squadron_Size* varied from 1 to 1000. *Colony size* after 3 years was used as output, averaged over 10 replicate simulations." [Becher et al. 2014]

(iii) The analysis technique employed, describing the details.

- *Example of the description of methods for global sensitivity analysis*:

"Since a full global sensitivity analysis was not computationally feasible, we followed a two-step protocol: (1) screening 72 selected model parameters to differentiate influential and non-influential parameters (remaining parameters were cast aside based on a first pre-analysis and results from previous sensitivity analyses described in Railsback et al. 2009; see TRACE document), and then (2) a variance-decomposition technique to identify, among the eight most influential parameters, those that reduce the output variance most when fixed to their "true" values. [...]

The screening step used an improved version of Morris's elementary effects method (Morris 1991; Campolongo et al. 2007). This method uses individually randomised one-factor-at-a-time designs to estimate the effects on model output of changes in parameter values; these effects are called elementary effects (EEs). The EEs are then statistically

analysed to measure their relative importance (see Thiele et al. 2014 for detailed description). We used the estimated mean of the distribution of the absolute values of the EEs, μ^*, as a sensitivity measure to establish the relative influence of each parameter. All 72 screened parameters were varied over five levels according to predefined ranges, the central value being the value used to calibrate the model (Table A19 in Section 7 of Appendix A). The number of tested settings was given by $r \times (k + 1)$, where r is the number of EEs computed per parameter and k the number of parameters. As we chose 50 EEs, this led to $50 \times (72 + 1) = 3650$ model runs." [Ayllón et al. 2016]

"We applied the variance-decomposition technique of Sobol (1993) to decompose the model outputs' variance into variances attributable to each input parameter while also evaluating the interaction between parameters. Sobol first-order sensitivity indices (S_i) measure the effect of varying a focus parameter alone but averaged over variations in other input parameters, thus providing information on the average reduction of output variance when the parameter is fixed. The total-effect indices (S_{Ti}) measure the contribution to the output variance of the focus parameter, including all variance caused by its interactions, of any order, with any other input parameters [...] The number of tested settings was given by $m \times (p + 2)$, where m is the size of the Monte Carlo sample matrix and p is the number of parameters to analyse." [Liukkonen et al. 2018]

- *Example of the description of methods for local sensitivity analysis*:
"To analyze the effect of variations in selected parameters on Fst values, we fitted linear regression models using the mean Global Fst values over generations 1,501- 2,000 (Fst values are already stabilized) as the response variable and the tested parameter as predictor; we fitted non-linear regression models when the R^2 of the linear model was lower than 0.70. We additionally fitted a multiple linear regression model to analyze the combined effect of tested parameters on Global Fst values, using linearized data (see equation in the Table 2)." [Baggio et al. 2018]

(iv) Settings of the other model components when this information is relevant.
- *Example of settings regarding input environmental time series*:
"Each simulation was run from the 1st of October of 1993 to the 30th of September of 2004 using the same environmental and hydraulic input used in model calibration." [Ayllón et al. 2016]

- *Example of settings regarding the spatial configuration:*
"Because the locations of haul out sites most likely influence the patterns, we simulated each of the parameter combinations keeping the same haul out site positions through all runs. The simulations were performed for 4.066 months as long as the movements of the calibration individual were monitored in 2011." [Liukkonen et al. 2018]

(v) Other technical details, such as number of replicates, and software used to implement the statistical algorithms.

"We used the *sensitivity* R package (Pujol et al., 2016), which implements the Monte Carlo estimation of the Sobol's indices using the improved formulas of Jansen (1999) and Saltelli et al. (2010)." [Liukkonen et al. 2018]

"For each parameter set, we ran 25 simulation replicates over 2,000 generations each without damming. Global Fst values were calculated every 10 generations […] Statistical analyses were performed in R version 3.3.3 (R Core Team, 2017)." [Baggio et al. 2018]

4 Simulation experiments:

1) Describe the aim of the experiment.

- *Example of experiments aimed at forecasting system dynamics*:

"We analyse the population's demographic and evolutionary dynamics under two simulation scenarios involving (1) warming resulting from climate change, and (2) climate change-induced warming plus stream flow reduction resulting from land use change, compared to (3) a baseline that includes the potential for evolutionary dynamics, but with no environmental change." [Ayllón et al. 2016]

- *Example of experiments aimed at selection of optimal management strategies*:

"We conducted a series of simulation experiments to assess whether and how Iberian pear seed arrival into the oldfield is influenced by (1) the density and distribution of planted trees and (2) by the preference of seed dispersers for aggregated vs. isolated fruiting trees. We used two typical tree densities or planting efforts (15 or 30 planted trees) and three tree distributions (aggregated, random, regular; Figure 2) to account for potential logistical and budgetary constraints of different restoration campaign designs (Rey-Benayas et al., 2008; Stanturf et al., 2014)." [Fedriani et al. 2018]

- *Example of experiments aimed at theory development/testing*:

"To evaluate which foraging decision rules best characterise the foraging behaviour of pink-footed geese we analysed whether each rule caused the model to reproduce a range of patterns observed in the field in 2005-2007 and 2011-2013 [...] We test the model's ability to reproduce the observed patterns (section 2.3) using five alternative foraging decision rules (FDRs) that differ in complexity." [Chudzinska et al. 2016]

"Theories are tested by implementing them in IBMs and determining how well they reproduce a variety of observed patterns, referred to as test patterns. By using test patterns observed at both the individual and population levels, we can identify theories that explain both individual behavior and the population-level responses that emerge from individual behavior. A priori, we identified eight test patterns of diel activity and habitat selection, many from the extensive laboratory and field experiments (cited below) of N. B. Metcalfe, N. H. C. Fraser, and colleagues at the University of Glasgow. Using the IBM, we reproduced the conditions under which the test patterns were observed, then observed whether IBM results reproduced the patterns." [Railsback et al. 2005]

- *Example of experiments aimed at understanding model behaviour*:

"We conducted various simulation experiments on different landscapes to illustrate and assess model behavior." [Carter et al. 2015]

2) Indicate the initial conditions and experiment settings, providing information about:
(i) Parameter values/ranges tested (with parameter space exploration algorithm if applicable), including external datasets, such as environmental time series or maps.
- *Example of experiments consisting in varying parameter values*:
"We varied growth conditions by simulating five values of the two reach-level parameters for food availability (drift concentration and production of benthic food), ranging from 33% to 300% of their baseline values. For survival conditions, we used five values of the parameters controlling daily survival probability of predation risk, from 98% to 102% of baseline values (corresponding to a ±50% range in the probability of surviving for 30 days)." [Railsback et al. 2014]

- *Example of experiments consisting in varying input environmental conditions*:
"Since temperature was shown to be the most important environmental factor affecting the population dynamics of *B. eunomia* (Radchuk et al., 2013a), we compared the predictions of dIBM and ySBM under a set of six temperature change scenarios. We used the predictions of Belgian National climate commission (Hoyaux et al., 2010) for Belgium to implement three mean change scenarios (low, moderate and high) covering the range of plausible change in the mean temperature over the next 100 years (increase in summer temperature by +2.4 to +7.2 ºC, increase in winter temperature by +1.4 to +4.4 ºC, see Appendix for details). Mean change scenarios were implemented by changing (i) in dIBM the mean of monthly temperature distributions as predicted for Belgium (Hoyaux et al., 2010); (ii) in ySBM the mean of distributions used for each life stage of the species (as this was the finest grain possible due to the model structure)." [Radchuk et al. 2014]

- *Example of experiments consisting in varying input spatial configuration*:
"We explored the effects of land use by synthesizing landscapes to represent wide ranges of variation in amount and distribution of habitat types. Bird foraging was simulated in each synthetic landscape […] Effects of intact forest area were evaluated by running the model with five landscape scenarios varying in forest area from zero to twice the baseline value (0-20% of total area). Forest was assumed to replace, or be replaced by, coffee habitat in equal parts shade and sun. As the area of forest and coffee habitat varied, the number of coffee patches was adjusted to maintain a constant size." [Railsback and Johnson 2014]

(ii) Model structure (only if different from the default model description).
"To study the stocking impacts on wild brown trout, several modifications of the model structure were needed. More specifically, one submodel, two breeds, and five parameters (*stocking-coefficient*, *num-stocked*, *trout-spawning-prob*, *stocked-spawning-prob*, *hybrid-spawning-prob*) were added. Both spawning and survival submodels were adapted to include phenotypic and genetic differences between hatchery-reared and wild trout." [Frank and Baret 2013]

(iii) Simulation settings, including time step, simulation length, stop conditions, and number of replicates.
"For each combination of factors, we ran 100 simulations, with each replicate comprising 1,800 time steps or hours (i.e. 75 days × 24 hr). The time period modelled (75 days) represents the "dispersal season" from mid September to the end of November, when ripe *P. bourgaenea* fruits are available to dispersers (Fedriani et al., 2012)." [Fedriani et al. 2018]

3) Regime shifts, events, and other scheduled model forcing, if applicable, explaining:
(i) When and to what values parameter sets are changed.
"We also assessed how mortality processes, such as female starvation, male challenges, and infanticide, are density dependent in the model. We created a 125 × 128 landscape with prey biomass production per cell set to the midpoint (6.255) of the lower (2.05 kg) and upper limit (10.46 kg) in Chitwan. The model was initialized with 50 adult females and 20 adult males […] Mortality was deactivated over the first 4 years to get territories established and reach quasi-stationary (i.e., stable population size over time) population dynamics more rapidly. Once the population reached a quasi-stationary point after 200 time steps, 50% of the adult females and males were removed from the model, and then various mortality processes and total tiger population size were evaluated for the next 20 years." [Carter et al. 2015]

(ii) Whether particular events are scheduled.
"A submodel that simulates the introduction of hatchery individuals in the Lesse River was created. Each year at week 27 (i.e., the end of March, corresponding to the beginning of the fishing season) and during 10 years, a fixed number of stocked trout (*num-stocked* parameter) was introduced in stream L. This number was calculated once, at week 27 of year 1, as the product of the stocking coefficient by the number of wild trout in stream L. The value of *stocking-coefficient* was sequentially set to 0.50, 0.70 and 0.90 to reflect the fact that the river is moderately stocked." [Frank and Baret 2013]

(iii) Whether there are structural changes forced upon the model runtime at predefined times.
"In each simulation, the model is run over 2,000 generations without barriers and in time step 2,001 a dam is added downstream of the fourth tributary. As a consequence, the modeled river network is fragmented into 2 subnetworks with 4 tributaries each, tributaries C and D being adjacent to the dam in one subnetwork, tributaries E and F in the other. The model is subsequently run over 1,000 generations with the presence of the dam. We analyzed three dam permeability scenarios: non-permeable, asymmetrical permeability and symmetrical permeability. Sensitivity analyses showed that Fst values were relatively insensitive to permeability (p) values (see results section 3.1), so we chose a low level of permeability (5%). In the simulations with asymmetrical permeability, dispersion is only possible from subpopulations A, B, C and D (source subnetwork) to tributaries holding subpopulations E, F, G and H (recipient subnetwork) (Fig. 4). This asymmetrical dispersion could represent either the movement of larvae and adults moving through spillways and fish ladders in the downstream direction or alternatively the upstream migration of adults moving through a fish ladder. In the simulations with symmetrical permeability (both ways), all these modes of dispersion are possible." [Baggio et al. 2018]

4) Variables observed and statistical analyses.

"We analysed the effects of the different restoration strategies on five model outputs or response variables: number of fox-dispersed seeds arriving to the oldfield, number of badger-dispersed seeds arriving to the oldfield, number of oldfield cells receiving seeds from aggregated trees, number of oldfield cells receiving seeds from isolated trees, and total number of oldfied cells receiving seeds [...]

To estimate the relative importance of the three main factors (number of planted trees, distribution of planted trees, dispersers preferences of aggregated vs. isolated trees), we partitioned the total variation of each of the five response variables (see Section 2.5) by analyzing the variance components. To this end, we used the Mixed procedure of SAS (SAS Institute 2016) and, as required for variance partitioning, the three factors were considered as random effects." [Fedriani et al. 2018]

References

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to 'evaludation': a review of terminology and a practical approach. Ecological Modelling, 280, 117-128.

Ayllón, D., Railsback, S.F., Vincenzi, S., Groeneveld, J., Almodóvar, A., and Grimm, V. 2016. InSTREAM-Gen: Modelling eco-evolutionary dynamics of trout populations under anthropogenic environmental change. Ecological Modelling 326: 36-53.

Baggio, R.A., Araujo, S.B.L., Ayllón, D., and Boeger, W.A. 2018. Dams cause genetic homogenization in populations of fish that present homing behavior: Evidence from a demogenetic individual-based model. Ecological Modelling 384: 209-220.

Becher, M.A., Grimm, V., Thorbek, P., Horn, J., Kennedy, P.J., and Osborne, J.L. 2014. BEEHAVE: a systems model of honeybee colony dynamics and foraging to explore multifactorial causes of colony failure. Journal of Applied Ecology 51(2): 470-482.

Carter, N., Levin, S.A., Barlow, A., and Grimm, V. 2015. Modeling tiger population and territory dynamics using an agent-based approach. Ecological Modelling 312(0): 347-362.

Chudzinska, M., Ayllón, D., Madsen, J., and Nabe-Nielsen, J. 2016. Discriminating between possible foraging decisions using pattern-oriented modelling: The case of pink-footed geese in Mid-Norway during their spring migration. Ecological Modelling 320: 299-315.

Dick, D.D.C., and Ayllón, D. 2017. FloMan-MF: Floodplain Management for the Moor Frog - a simulation model for amphibian conservation in dynamic wetlands. Ecological Modelling 348: 110-124.

Fedriani, J.M., Wiegand, T., Ayllón, D., Palomares, F., Suárez-Esteban, A., and Grimm, V. 2018. Assisting seed dispersers to restore oldfields: An individual-based model of the interactions among badgers, foxes and Iberian pear trees. Journal of Applied Ecology 55(2): 600-611.

Frank, B.M., and Baret, P.V. 2013. Simulating brown trout demogenetics in a river/nursery brook system: The individual-based model DemGenTrout. Ecological Modelling 248: 184-202.

Liukkonen, L., Ayllón, D., Kunnasranta, M., Niemi, M., Nabe-Nielsen, J., Grimm, V., and Nyman, A.-M. 2018. Modelling movements of Saimaa ringed seals using an individual-based approach. Ecological Modelling 368: 321-335.

Lorscheid, I., Heine, B. O., & Meyer, M. (2012). Opening the 'black box'of simulations: increased transparency and effective communication through the systematic design of experiments. Computational and Mathematical Organization Theory, 18(1), 22-62.

Phang, S.C., Stillman, R.A., Cucherousset, J., Britton, J.R., Roberts, D., Beaumont, W.R.C., and Gozlan, R.E. 2016. FishMORPH - An agent-based model to predict salmonid growth and distribution responses under natural and low flows. Scientific Reports 6: 29414.

Piou, C., and Prévost, E. 2012. A demo-genetic individual-based model for Atlantic salmon populations: Model structure, parameterization and sensitivity. Ecological Modelling 231: 37-52.

Radchuk, V., Ims, R. A., & Andreassen, H. P. 2016. From individuals to population cycles: the role of extrinsic and intrinsic factors in rodent populations. Ecology, 97(3), 720-732.

Radchuk, V., Johst, K., Groeneveld, J., Turlure, C., Grimm, V., and Schtickzelle, N. 2014. Appropriate resolution in time and model structure for population viability analysis: Insights from a butterfly metapopulation. Biological Conservation 169(0): 345-354.

Railsback, S.F., and Harvey, B.C. 2002. Analysis of habitat-selection rules using an individual-based model. Ecology 83(7): 1817-1830.

Railsback, S.F., and Johnson, M.D. 2011. Pattern-oriented modeling of bird foraging and pest control in coffee farms. Ecological Modelling 222(18): 3305-3319.

Railsback, S.F., and Johnson, M.D. 2014. Effects of land use on bird populations and pest control services on coffee farms. Proceedings of the National Academy of Sciences 111(16): 6109-6114.

Railsback, S.F., Harvey, B.C., and White, J.L. 2014. Facultative anadromy in salmonids: linking habitat, individual life history decisions, and population-level consequences. Canadian Journal of Fisheries and Aquatic Sciences 71(8): 1270-1278.

Railsback, S.F., Harvey, B.C., Hayse, J.W., and LaGory, K.E. 2005. Tests of theory for diel variation in salmonid feeding activity and habitat use. Ecology 86(4): 947-959.

Railsback, S.F., Harvey, B.C., Jackson, S.K., and Lamberson, R.H. 2009. InSTREAM: the individual-based stream trout research and environmental assessment model., U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station, Albany, CA.

Thiele, J. C., Kurth, W., Grimm, V. 2014. Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. Journal of Artificial Societies and Social Simulation, 17(3), 11.