To what extent does incorporating more complexity into models make them less informative?

Kamran Pentland

EPSRC & MRC Centre for Doctoral Training in Mathematics for Real-World Systems, University of Warwick

Abstract

In this report, the effects of increased realism and complexity in various scientific models is investigated. It is well established that methods and techniques that incorporate too many features and parameters risk producing explainable results, creating over-fit models and often unnecessarily waste computational resources. On the other hand, simplistic methods are often viewed as too idealistic for the real-world, sometimes thought to produce less accurate results than complex methods. By reducing the number of idealistic assumptions a model makes, do results necessarily become more accurate and informative? Short model examples, case studies and experiments are discussed, with regard to the informativeness of the results they produce, using both simplistic and more complex modelling techniques. The perception that more complex methods are inherently 'better' and more accurate than more simplistic techniques are also discussed. It is concluded that both simple and complex methods must be selected with care, depending on what one wishes to discern from a model, the resources available and the informativeness of the results that it produces.

1. Introduction

Broadly speaking, scientific models can be defined as "idealisations, abstractions or analogies" [9, p.1] that aid the understanding of our physical, social and technological realities. In essence they are simplifications, based on fundamental assumptions, of a system we wish to try to understand. Their specific definitions, however, depend enormously on the modeller's objectives and the field under study (see [9, p.55]). Whereas experimental models tend to consist of physical setups in lab-based environments (think wave machines or the Large Hadron Collider), others are often purely analytical, comprising of theory, flow charts, mathematical equations and so forth. The invention of the computer revolutionised modelling capabilities, allowing expensive physical experiments to be replaced with cheaper simulation-based models. Such problems that have vast numbers of parameters and outcomes; extremely large data sets; and no closed-form analytic solutions[1] in areas from epidemiology and statistics to operational research and genomics [1, 12] can now be investigated on drastically smaller timescales.

Models underpin almost all scientific thinking, they are tools primarily used to *understand* the way in which a complex system functions, not necessarily to predict and forecast the evolution of such systems (although this knowledge is certainly beneficial to have!). If constructed correctly, they can provide invaluable theory, information, visualisations, and, very rarely, exact solutions of complex systems. Modelling objectives differ depending on the requirements of the modeller. Whereas scientific objectives might be a combination of the aforementioned features, depending on research goals and funding, governmental bodies often seek more predictive (forecast) models to inform policy[2]. This contrasts with industry modellers that may seek to automate and optimise manufacturing processes.

The features that make models more realistic, and therefore usually more complex, can include anything from a system's physical scale (micro, meso, macro etc.); its number of (non-)physical dimensions and parameters; nonlinear effects and interactions; feedback loops; collective behaviour; or bifurcating parameters, to name but a few [17]. That is, to incorporate additional features, rules or observations in order to improve how a model reflects or simulates our physical or social reality, according to humanistic perceptions.

In this short report we examine the effects of introducing more realistic features and data into models and how significantly this impacts their informativeness. Models incorporating far too much realism risk over-fitting, wasting computational resources and becoming incoherent to non-technical individuals whilst simpler models can often be viewed as too idealistic or basic for real-world needs. We investigate this trade-off and discuss the benefits and weaknesses of a simpler versus a more complex approach to modelling and why complex models are often favoured over more simplistic ones through the use of informative examples, case studies and experiments.

[1] See the Navier-Stokes equations for an example of a fluid mechanics problem that currently has no closed-form analytic solutions in three-dimensional space but can however be approximated extremely well using numerical simulations (note that there is a small prize if you can show these solutions exist!) [18].

[2] For example, the vast majority of the UK government's initial response efforts to stop the spread of COVID-19 largely focused on forecast models provided by the Imperial College COVID-19 Response Team among others [8].

2. Idealism vs. Realism

Depending on the system in question, the distinction between a simplistic and complex model can be difficult, and not a strict dichotomy. In an effort to contrast and compare them however, we label them opposites, where a simple model is one that is intuitive, understandable or explainable to individuals with limited knowledge of the problem under investigation. This, however, does not mean they are 'easy' or 'fully solved' systems. Complex models, contrary to discussions in [10], *can* be thought of as some unquantifiable function of, not just model inputs, but also model outputs; behaviour; understandability; comprehensibility; and the ease of computation or construction of said model. Attempts to classify complexity in general and specific fields have been made, however it is not a straight forward task [13].

Regardless of whether a model is exploratory, analogical, phenomenological or some other type [14], they can be described qualitatively, if not quantitatively, by their level of realism compared to our physical reality (see Figure 1). Idealisation removes features that are not essential to understand the basic properties of a system (i.e. assuming individuals behave rationally, in social interaction models) whereas increased realism comprises the opposite, adding more features to a model in order to reflect our own reality more accurately.

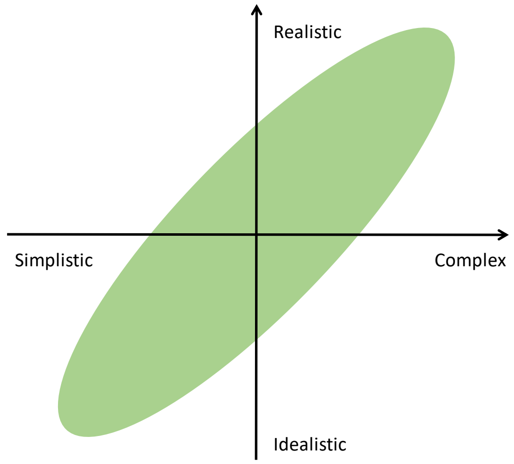


Figure 1: A simplistic categorisation diagram for a generic model on a scale of idealism and realism versus simplicity and complexity.

By reducing model idealism, we expect model complexity to increase, hence most models exist somewhere in or around the green area in Figure 1. Think of the Large Hadron Collider existing toward the upper right area and idealistic models of perfect competition in economic theory toward the lower left. Clearly not all models fit this (very basic) taxonomy[3] and attempting to categorised every type of model would stir up vigorous debate and criticism among researchers. Figure 1 provides a simple, overly generalised, measure for model simplicity versus complexity. Incorporating our interest, model informativeness, onto a third axis would be perfect for this report, however this sort of classification is far beyond any reach. Instead we now consider a simple idealistic model and increase its realism in order to qualitatively assess its informativeness.

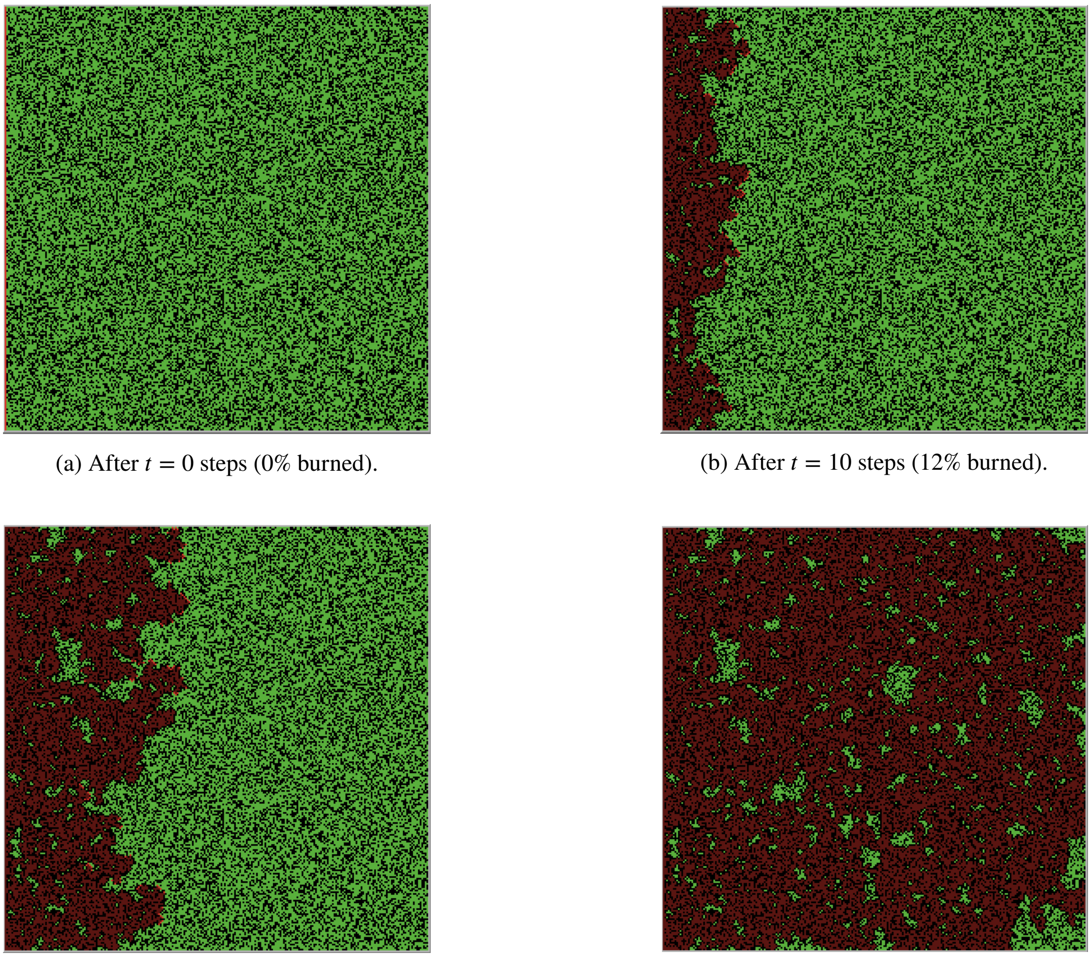
3. Increasing realism: An agent-based example

We examine a simplistic agent-based model (ABM), developed in NetLogo [19, 20, 21], that simulates a forest fire under some idealising assumptions (see Appendix A for further details). It models the spread of an initial line of burning trees, beginning on the western boundary of a two-dimensional grid, that propagates eastward, see Figure (2a)-(2d) for visualisations. An initial density of trees is specified by the user whilst a single output is measured: the percentage of initial trees burned once the fire burns out.

Upon running an 'experiment' in NetLogo (see Appendix A), we can determine that some form of nonlinear relationship exists between the percentage of forest burned and the initial tree density, see Figure 3. By increasing the number of simulations for each density, an empirical range can be determined, dependent on the stochastic (random) nature of the initial tree density (i.e. dense areas of forest are more likely to burn down, than less dense areas). This simple model allows us to visualise the spatial spread of the fire however it is limited by assuming trees are homogeneous; that fire spreads with probability one if they are adjacent (horizontally and vertically); no external factors exist (wind, rain etc.); and that the grid itself is not representative of real forest structure.

Clearly this simple model cannot replicate the true spread of a forest fire so let us add some more realistic effects in order to try to improve or discern more information from the results. Adjusting some code in NetLogo, we allow the fire to spread diagonally to other tiles. If the forest was partially wet and contained varying tree types, not every tree would (or could) catch fire, so now we include a varying probability p that a burning tree ignites its neighbour. A toggle can also be adjusted to include sparks jumping over empty tiles and igniting other trees. The ensemble averages for different values of p, with and without spark jumps on or off, are given in Figure 4. The first obvious observation is that the spark jumps have negligible impact on the results. The second, and most dramatic, is the downward shift in critical density

[3] Social simulation models have been studied and are known to be incredibly complex whilst simultaneously being relatively idealistic in their underlying assumptions [5]. Clearly these sorts of models would sit outside the green area in Figure 1.



(a) After $t = 0$ steps (0% burned).

(b) After $t = 10$ steps (12% burned).

(c) After $t = 20$ steps (33% burned).

(d) After $t = 32$ steps (81% burned).

Figure 2: Visualisations of the agent-based forest fire model from NetLogo as it evolves over time from (a)-(d) with an initialised tree density of 62%.

(almost 20%) caused by allowing the fire to spread in eight directions instead of four. Third is that by decreasing p, we observe that, a higher initial density is required for the same amount of forest to burn (on average). This increase in initial density varies, again, in some nonlinear manner that we cannot discern from this chart and hence further investigation is required.

We discover that by incorporating more realistic features into the simple ABM, they do not significantly improve our understanding of the nonlinear relationship between initial density and percentage burned. At a significant cost in computational effort[4] we did however learn of another (unknown) nonlinear interaction between p and initial tree density. Although limited analysis exposes this relationship, it comes at high cost as the extra features inadvertently generate approximately 50,000,000 possible parameter combinations (more with the random initial condition) for the model. Not all combinations could, or necessarily should be, explored within limited time however.

We conclude by saying that the simple model, and to some extent the complex one, are excellent tools for visualising, teaching and demonstrating how an idealised forest fire might propagate in a dense forest. However it becomes clear that even if the model had perfectly fitted parameters; a scaled up forest of interest; more realistic tree structures and locations; as well as gaps for roads and towns, it would certainly be computationally infeasible to simulate and analyse in good time. The additional realism generated results that were more difficult to analyse, whilst being only slightly more, if not equally, informative than the simpler ABM. Whilst this exploratory ABM lacks predictive power that some groups require to predict the spread of a fire, [11] suggests that the visualisation aspect can however be used as a tool for testing firefighting and suppression tactics under real-world conditions.

Therefore, if resources are available, the additional complexity may be worth the effort[5] and hence researchers need to determine, on a case-by-case basis, whether additional re-

[4] Each curve plotted in Figure 4 took an average of forty minutes to simulate.

[5] Note that the addition of wind was another feature of the model not included here for simplicity but available on NetLogo.

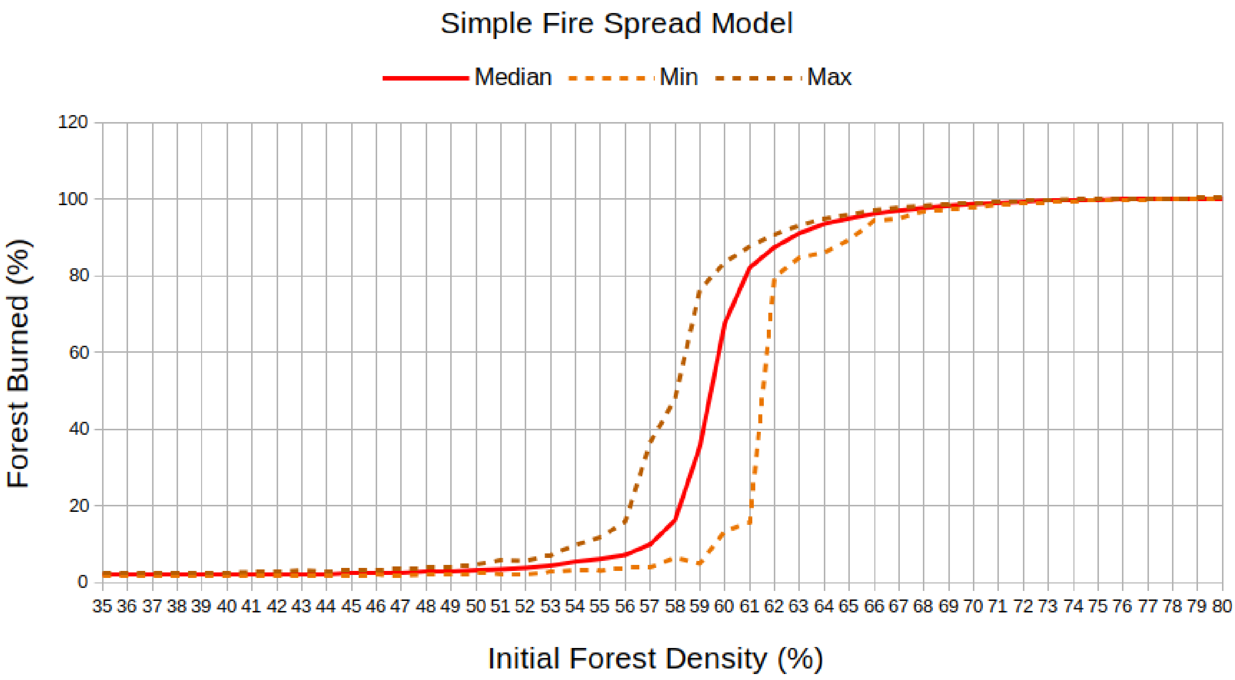


Figure 3: The relationship between initial forest density and the percentage forest burned, once the fire has gone out. For each density, 100 simulations were run and the plot gives the range and median percentage of forest burned.

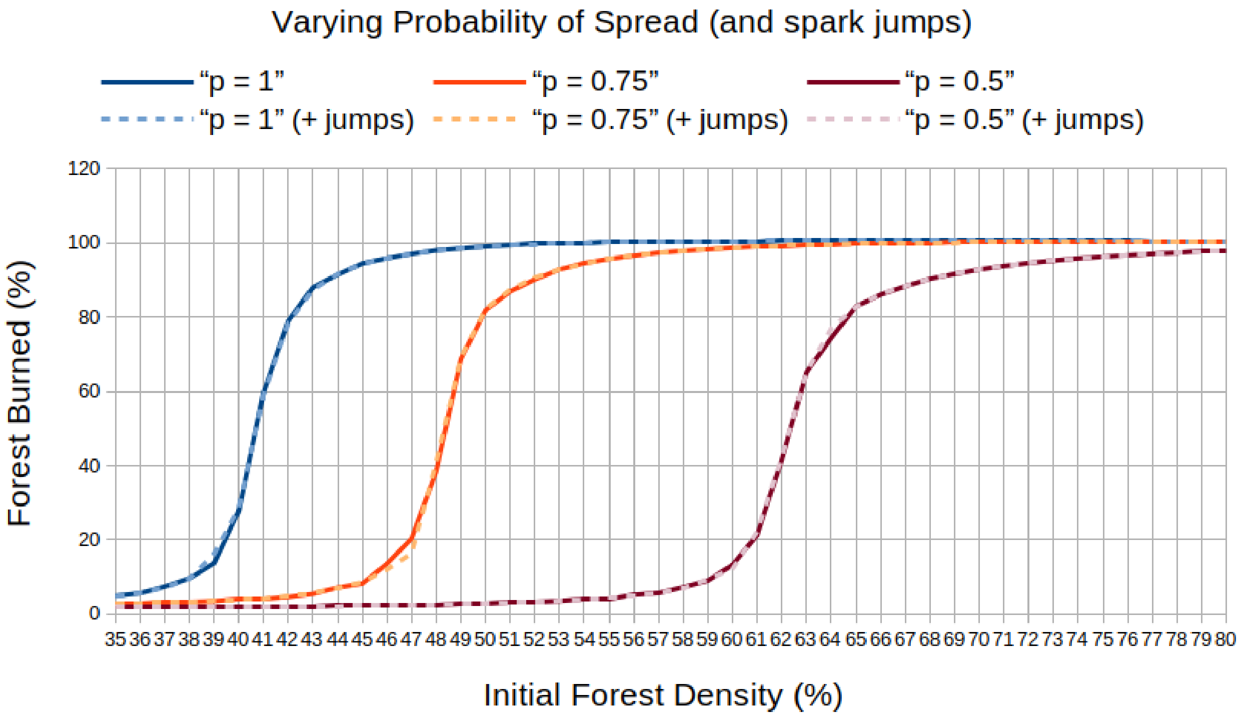


Figure 4: The varying relationship between initial density and forest burned is shown for various p values. Note that in these scenarios fire was also allowed to spread in eight directions instead of the original four.

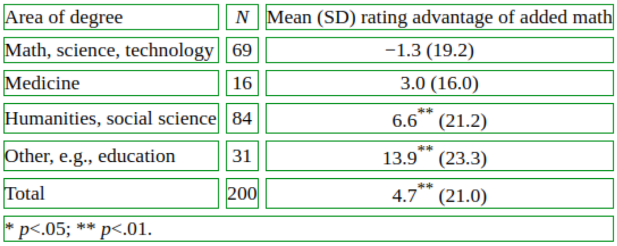
alism (not necessarily in this small model) is worth the time, money and effort for the results they provide.

4. Perceptions of complex models

Moving away from a specific model, we now discuss how complex models are perceived and approached by different individuals and the impact this might have on their informativeness. As technology advances and we move into a more data heavy world, UK researchers and funding bodies[6] are favouring the development of more complex modelling techniques in order to tackle today's most challenging problems in fields such as AI, meteorology and epidemiology, with the aim of developing models with that mimic and forecast real world phenomenon. Numerical weather forecasts [2] that require solving vast numbers of coupled nonlinear equations with huge data sets on supercomputers are examples of highly complex models that require increased realism in order to provide accurate results. This demonstrates, in some areas, additional realism or complexity is a *requirement*, based on necessity, rather than the preference of the modeller. Such realistic methods have begun to be incorporated into more advanced forest fire models that clearly require a specific level of realism [4].

In these cases, increased realism improves model informativeness drastically, however this may not always be the case and, mistakenly, complex models can often be unquestionably perceived as 'better' or more informative than simplistic alternatives without good reason, explanation or validation. One review [16], quantitatively summarises a set of papers on population projection models and states that using complex statistical methods (ARIMA[7]) makes projections no more accurate than simple (linear/exponential) approaches. It was pointed out that the complex methods were only more accurate when forecasts were made for small population towns and cities, not on larger scales.

Another study discusses the "Dr. Fox Phenomenon" [3] with regard to unintelligible writing in management science journals. It was found that a positive correlation existed between a journal's academic prestige and its "fog index" (how readable the journal was), suggesting that harder to read material was viewed as being of a higher quality, regardless of the content. More recently, [7] discusses the inclusion of nonsensical mathematical equations into research abstracts and how researchers (in non-mathematical fields) judged them to be of higher quality than those without. Figure 5 summarises the results of the study and show how much higher ratings were awarded by participants with less mathematical training.



Area of degree	*N*	Mean (SD) rating advantage of added math
Math, science, technology	69	−1.3 (19.2)
Medicine	16	3.0 (16.0)
Humanities, social science	84	6.6^{**} (21.2)
Other, e.g., education	31	13.9^{**} (23.3)
Total	200	4.7^{**} (21.0)
* *p*<.05; ** *p*<.01.		

Figure 5: A reproduction of Table 1 from 5 displaying the mean ratings of a research abstract, that included a nonsensical mathematical equation, by participants from participants with a variety of undergraduate degrees.

These studies beg the question as to whether we perceive and favour difficulty or complexity as a prerequisite for a high quality model, regardless of its usefulness and informativeness. Simply because a model's development and its results were more difficult to obtain, are they necessarily better than those obtained using a much simpler method? This may seem unlikely as complex models are often based on their simpler counterparts, however, as demonstrated by the forest fire and population projection results, increasing complexity often amounts to similar, maybe slightly more detailed, results than their simplistic counterparts for much more work. Whilst not definitive throughout science, one must ask if perhaps *some* models are built for the sake of complexity and whether, in order to reduce the risk of being overlooked for future funding or higher ranking journal publications, some research groups feel obliged to use cutting edge methods when a simpler one might do.

5. Conclusions

We observed, even with a simple ABM, how increased realism introduced nonlinearity and a measure of obscurity into model results, making them tougher to explain and more time consuming to obtain without necessarily improving their informativeness to much extent. This exploratory model was however naturally insightful, not because it perfectly reproduced forest fire spread but because it provided essential information about the relationship between initial density and trees burned. The population projection methods also demonstrate that blindly applying more advanced forecast methods did not yield more accurate results than simpler techniques, advocating the use of Occam's razor when given a wide choice of modelling techniques. We have also seen that additional complexity can sometimes be misinterpreted as a prerequisite for obtaining a superior model without true justification.

Simplistic models provide excellent tools for exploratory research and teaching, often having applications in multiple fields of study (i.e. Nash equilibria in economic, evolutionary and computer game theory), an attribute much more difficult to achieve with more complex case-specific models.

[6] Almost all of the 75 Centres for Doctoral Training (funded by EPSRC in 2019) in the UK focus on some form of complex modelling in the physical sciences [6].

[7] These are Auto Regressive Integrated Moving Average models used for fitting historical data and forecasting future data values.

Complex models can often behave like ‘black box’ machines and it is much easier to discover faults and flaws in simplistic models. Simpler methods can provide insightful, often preliminary, results much more easily than complex versions, however individuals who require predictory models, in applications from weather to infectious disease modelling, may be forced to use more challenging techniques. In these cases and where simplistic methods have been exhausted, more complex methods may be unavoidable, in which case the modeller faces tough choices about how to optimise parameters and minimise computation time, errors inherent in the model and other associated costs. If we wish to understand a new problem or system from scratch, selecting the most important features and creating a simple model is going to be the best place to start. This may seem obvious, however as we have seen with the population projection methods, modellers sometimes opt for the more complex technique when in fact a simpler method may do the same job.

Recalling Figure 1, we conclude that an optimal balance must be struck between what a modeller wishes to discern from a model, the level of realism required to produce meaningful and informative results and the complexity that can be managed. Clearly we strive for accurate models that replicate real world phenomenon but not when it jeopardises the reliability and informativeness of results. Deichsel and Pyka [15] make an interesting point that "we make models in order to reduce the complexity of the real world, not to mirror it." Whilst I agree with this philosophy, I do however believe that the overwhelming curiosity and intellectual challenge of creating models ceases to exist if a model does not at least try to mirror a little of our own reality.

Appendix A

The forest fire model is run in NetLogo using ‘Fire Simple Extension 3 Model’, built originally by [20], adapted slightly for the purposes of these experiments. The two-dimensional grid consists of small tiles which either contain nothing (black), a tree (green), a tree on fire (red) or a burnt out tree (dark brown). The user specifies a single input, the initial tree density, which places trees randomly on tiles throughout the grid (see Figure 2a). Upon simulation, the fire propagates on the visualiser over time, see Figures (2b)-(2d), whilst a single output is measured: the percentage of initial trees that have burned. The initial governing rules (adapted later) of the ABM are that:

1. A burning tree can only pass fire to its horizontal and vertical neighbours (north, east, south and west) in one step before then burning out itself (it cannot ignite others once burnt out).
2. A burning tree passes on the fire with probability $p = 1$.
3. Fires cannot jump over empty or burnt out tiles.

Initial Density of Forest (35-80%)			
All 8 directions?	p value?	Sparks?	Wind?
No	$p = 1$	No	Unused
Yes	$p = 1$	No	Unused
Yes	$p = 0.75$	No	Unused
Yes	$p = 0.5$	No	Unused
Yes	$p = 1$	Yes	Unused
Yes	$p = 0.75$	Yes	Unused
Yes	$p = 0.5$	Yes	Unused

Table 1
A list of the seven BehaviourSpace experiments run in NetLogo for the forest fire model over 46 initial densities and 100 individual repetitions.

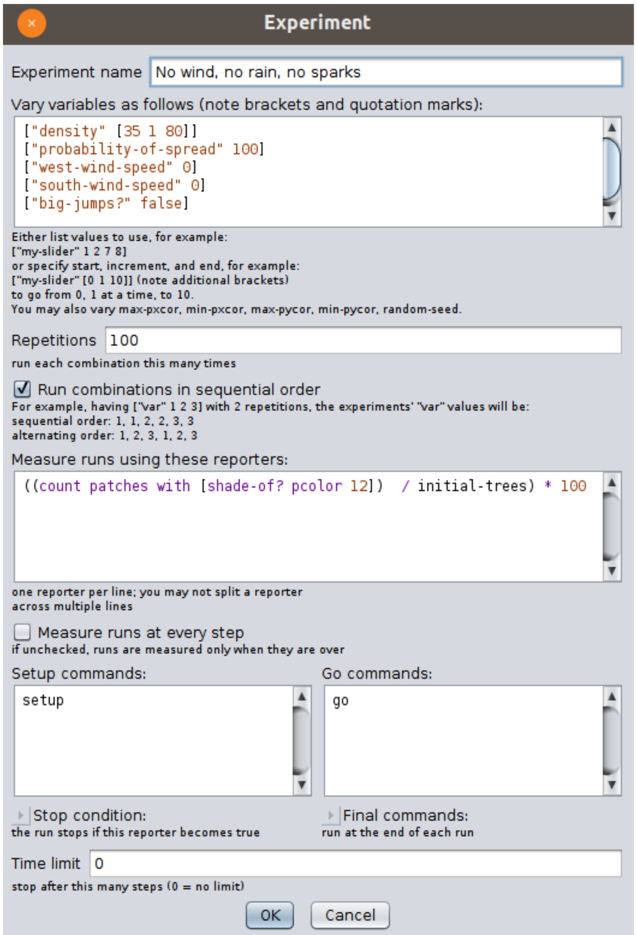
The BehaviourSpace[8] in NetLogo is an excellent tool that allows us to run experiments on the model by varying its parameters systematically over a repeated number of simulations. Clearly not all parameter values in the parameter space can be examined in good time and neither is this necessary. In total there would have been around 100 x 100 x 50 x 50 x 2 = 50,000,000 different combinations of parameters in the parameter space, not including the random initial placement of trees which increases this number further.

Clearly it is infeasible to run all possible scenarios and hence we examine a certain subset to identify the system’s overall behaviour. For example, it was found that below 35% and above 80% initial density, the percentage of trees burned always reached 0% or 100% as the forest was too sparse or too dense. This automatically cuts the number of combinations down by over half. The incorporation of wind was also unnecessary for determining percentage forest burned as it only affected the spatial aspect (the parts) of the forest that burned in each simulation, not the total percentage.

A list of the different experiments with varying parameters is given in Table 1. Each of these simulations was run 100 times on four computer processors, in parallel, in order to generate an ensemble average of the percentage of trees burned for different initial tree densities (stopping once the fire goes out). This helped to remove the stochastic fluctuations generated by the random initial conditions. Each experiment in Table 1 was equivalent to running a single simulation 4600 times (46 different initial densities times 100 simulations). This required around forty minutes of computation time to run, hence it becomes easy to understand how exploring ABM parameter spaces becomes very expensive and time consuming even for such a simple model.

A picture showing the layout of a BehaviourSpace experiment that initialises a forest fire with $p = 1$ probability of spread, zero wind and no sparks over 35-80% initial densities is given in Figure 6. It details how to enter values, how many repetitions are required (100), if runs are measured at

[8] See https://ccl.northwestern.edu/netlogo/docs/behaviorspace.html for full details on BehaviourSpace in NetLogo



Experiment name: No wind, no rain, no sparks

Vary variables as follows (note brackets and quotation marks):

```
["density" [35 1 80]]
["probability-of-spread" 100]
["west-wind-speed" 0]
["south-wind-speed" 0]
["big-jumps?" false]
```

Either list values to use, for example:
["my-slider" 1 2 7 8]
or specify start, increment, and end, for example:
["my-slider" [0 1 10]] (note additional brackets)
to go from 0, 1 at a time, to 10.
You may also vary max-pxcor, min-pxcor, max-pycor, min-pycor, random-seed.

Repetitions: 100
run each combination this many times

☑ Run combinations in sequential order
For example, having ["var" 1 2 3] with 2 repetitions, the experiments' "var" values will be:
sequential order: 1, 1, 2, 2, 3, 3
alternating order: 1, 2, 3, 1, 2, 3

Measure runs using these reporters:

```
((count patches with [shade-of? pcolor 12])  / initial-trees) * 100
```

one reporter per line; you may not split a reporter
across multiple lines

☐ Measure runs at every step
if unchecked, runs are measured only when they are over

Setup commands:

```
setup
```

Go commands:

```
go
```

Stop condition:
the run stops if this reporter becomes true

Final commands:
run at the end of each run

Time limit: 0
stop after this many steps (0 = no limit)

Figure 6: A screenshot showing how the first experiment in Table 1 is coded into NetLogo's BehaviourSpace. Note that to adapt the direction of fire spread from 4 to 8, a code is changed in the model outside of BehaviourSpace.

each time step or just the final step and the quantity being measured by the experiment. In our case it counts the number of brown (burnt) patches divided by the initial number of green (trees) and multiplies by 100 to obtain a percentage. Further options ensure results can be output to a .csv file and that repetitions can be run in parallel on multiple computer cores.

References

[1] David Adam. *Special report: The simulations driving the world's response to COVID-19*. URL: https://www.nature.com/articles/d41586-020-01003-6. (accessed: 08.04.2020).

[2] Klaus Weickmann et al. *The Use of Ensemble Forecasts to Produce Improved Medium Range (3-15 days) Weather Forecasts*. URL: https://web.archive.org/web/20100528082602/http://www.esrl.noaa.gov/psd/spotlight/12012001/. (accessed: 08.05.2020).

[3] J. Scott Armstrong. "Unintelligible Management Research and Academic Prestige". In: *Interfaces* 10.2 (Apr. 1980), pp. 80–86. ISSN: 0092-2102. DOI: 10.1287/inte.10.2.80.

[4] Terry L. Clark et al. "A Coupled Atmosphere Fire Model: Convective Feedback on Fire-Line Dynamics". In: *American Meteorological Society* (June 1996). DOI: 10.1175/1520-0450(1996)035<0875:ACAMCF>2.0.CO;2.

[5] Paul Davidsson. "Agent Based Social Simulation: A Computer Science View". In: (2002).

[6] EPSRC. *Seventy five Centres for Doctoral Training announced by UKRI*. URL: https://epsrc.ukri.org/newsevents/news/seventy-five-centres-for-doctoral-training-announced-by-ukri-to-develop-the-skills-needed-for-uk-prosperity/. (accessed: 29.04.2020).

[7] Kimmo Eriksson. "The nonsense math effect". In: *Judgment and Decision Making* 7.6 (2012), pp. 746–749. URL: http://intercult.su.se/publications/2012/Eriksson%7B%5C_%7D2012%7B%5C_%7DJDM.pdf.

[8] Neil M Ferguson et al. *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*. Tech. rep. March. Imperial College London, 2020, pp. 3–20. DOI: 10.25561/77482. URL: https://doi.org/10.25561/77482.

[9] Philip Gerlee and Torbjörn Lundh. *Scientific models: Red atoms, white lies and black boxes in a yellow book*. Springer International Publishing, Jan. 2016, pp. 1–96. ISBN: 9783319270814. DOI: 10.1007/978-3-319-27081-4.

[10] Kesten C. Green and J. Scott Armstrong. "Simple versus complex forecasting: The evidence". In: *Journal of Business Research* 68.8 (Aug. 2015), pp. 1678–1685. ISSN: 01482963. DOI: 10.1016/j.jbusres.2015.03.026.

[11] Xiaolin Hu and Yi Sun. "Agent-based modeling and simulation of wildland fire suppression". In: *Proceedings - Winter Simulation Conference* (2007), pp. 1275–1283. ISSN: 08917736. DOI: 10.1109/WSC.2007.4419732.

[12] Eric S. Lander et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (Feb. 2001), pp. 860–921. ISSN: 00280836. DOI: 10.1038/35057062.

[13] Steven M. Manson. "Simplifying complexity: A review of complexity theory". In: *Geoforum* 32.3 (Aug. 2001), pp. 405–414. ISSN: 00167185. DOI: 10.1016/S0016-7185(00)00035-X.

[14] Stephan Hartmann Roman Frigg. *Models in Science*. URL: https://plato.stanford.edu/entries/models-science/#ModeRealLawsNatu. (accessed: 07.04.2020).

[15] Simon Deichsel and Andreas Pyka. "A Pragmatic Reading of Friedman's Methodological Essay and What It Tells Us for the Discussion of ABMs". In: *Journal of Artificial Societies and Social Simulation* (Oct. 2009).

[16] Stanley K. Smith. "Further thoughts on simplicity and complexity in population projection models". In: *International Journal of Forecasting* (1997). ISSN: 01692070. DOI: 10.1016/S0169-2070(97)00029-0.

[17] Stephen Strogatz. *Nonlinear Dynamics And Chaos: With Applications To Physics, Biology, Chemistry, And Engineering (Studies in Nonlinearity)*. Westview Press, 2001.

[18] Wikipedia. *Millennium Prize Problems*. URL: https://en.wikipedia.org/wiki/Millennium_Prize_Problems. (Accessed: 08.04.2020).

[19] Uri Wilensky. *NetLogo*. Evanston, IL, 1999.

[20] Uri Wilensky. *NetLogo Fire Simple Extension 3 model*. Evanston, IL, 2006. URL: http://ccl.northwestern.edu/netlogo/models/FireSimpleExtension3.

[21] Uri Wilensky and William Rand. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NETLogo*. Cambridge, MA: MIT Press, 2015. ISBN: 9780262731898. URL: http://ccl.northwestern.edu/netlogo/..