Why the term prediction is overused

Judith A. Verstegen[0000−0002−9082−4323] and Simon Scheider[0000−0002−2267−4810]

Department of Human Geography and Spatial Planning, Utrecht University, Utrecht, The Netherlands, j.a.verstegen@uu.nl, s.scheider@uu.nl

Abstract. While a model prediction is a probabilistic claim about a system state to transpire in the future, a model projection is an if-then statement about the potential future of a system, by definition subject to (changes in) boundary conditions with an unknown likelihood. Despite a robust body of literature on the various potential purposes of models - and to predict is only one of these purposes - some modellers tend to refer to all their model outputs as predictions, while they are more often projections or neither of these two. Both geosimulation and spatial machine learning scholars are careless in how they refer to their model outputs. This is confusing for all involved and especially for the general public, for whom the model output is usually the only model component they get to see. In this paper we provide definitions, justifications, and a decision tree for classifying model outputs. This can help the GIScience community to gain clarity about what their model output entails.

Keywords: geosimulation · geoAI · purpose · model validity.

DOI: https://doi.org/10.25436/E2MK55

1 Introduction

As a geosimulation modellers, the question most often asked to us is if we really believe in our models' predictions. The question comes from a wide range of people, colleagues from our own department, reviewers, peers at conferences, and the general public at events. This is somewhat surprising given the fact that we ourselves never ever use the term *prediction* in presentations or papers, but others do [1–3].

Instead, we use the term *projection* for model results, see, e.g., [4]. This may seem like an arbitrary difference, but it is not. A projection is a description of the future, but not one we are claiming to become reality. Making this explicit is crucial for managing the expectations of the direct users of model outputs as well as the general public.

In this manuscript, we clarify the distinction between a prediction and a projection, and cases where model output should be classified as neither of these. Furthermore, we highlight the differences in (as well as the wrong) usages of these terms in the geosimulation and spatial machine learning communities, present

arguments why precise terminology is important to resolve these issues, and link them to existing past and ongoing debates in GIScience.

2 Definitions

In the physical sciences, and especially in the climate science domain, scientists usually differentiate between predictions and projections. We follow the definitions applied in that domain [5]:

- A **prediction** is a probabilistic claim that something will happen in the future based on knowledge of the current state, i.e., what we *expect to happen* with a specified probability. A prediction assumes that future boundary conditions are known (same as now or changes predicted with confidence) or changes in them have no significant influence.
- A **projection**, in contrast, is an "if-then" (or conditional) statement, for example to evaluate the effects of different potential interventions in the system. Thereby, it specifically controls for changes in the system's boundary conditions in an experimental fashion, but does not assume any likelihood of these changes to occur.

Thus, a projection is a *conditional statement* that something may happen in the future if certain boundary conditions develop. At the same time, we are usually fairly certain that these conditions will *not* develop unless someone takes action, e.g. policy makers. A projection is not a prediction because it intentionally does not aim (only) at the most probable conditions, but also at other conditions with a low and/or unquantifyable probability to transpire.

The definitions above are not merely an epistemological argument, but also used by major organizations in the climate dialogue. For example, the Intergovernmental Panel on Climate Change (IPCC) provides clear definitions of the terms in their reports between predictions and projections and is relatively consistent in their usage [6].

MacCracken [5] distinguishes two other terms: forecast and scenario. The first is closely related to a prediction, but the "best guess" instead of a probabilistic claim, while the second is closely related to a projection but an indication of possibilities, rather than probabilities. We decided to stick to only prediction and projection in this paper because: 1) it may be hard enough to let the community adopt one new term, 2) the distinction with the other two terms is not obvious, and 3) the discourse on what a scenario entails in different scientific domains that employ models can easily fill a separate paper (e.g. [7]).

3 Geosimulation

> "The power of modelling comes from making an informal set of ideas formal" [8]

Geosimulation models simulate interactions between humans or animals and their environment with two main modelling paradigms: spatial agent-based models and cellular automata. They are computational formalizations of domain-specific spatial-process or behavioral theories[1], e.g., from ecology, hydrology, or spatial planning. That is, they are not algorithms trained on data but an explicit implementation of our knowledge about a spatial system. Geosimulation models are not exclusively meant to predict something; scholars have recognized more than a decade ago that purposes of models vary between studies [9]; purposes named are to predict, to explain [9], to research [12], to inform management decisions [10, 12], to describe, to explore, to illustrate, to draw an analogy, to interact [8], facilitate discussion amongst interdisciplinary research teams and stakeholders, to formalize assumptions, and to act as a repository for data [11]. In line with the recognition of different purposes, ecological modellers [12] and later the agent-based modelling community [13] have redefined the concept of model validity from "corresponds to the real system" to "is adequate for its intended purpose".

Yet, this discussion has not been transferred to how we refer to our model outputs; the terms introduced in the previous section are used inconsistently [5, 13], although some put disclaimers in their text that the term prediction should not be seen in a strict "statistical/econometric sense" [1]. But why not be strict? It would be relatively easy to connect the nouns used for model output to the already defined verbs for model purposes. Out of the potential model purposes provided above, the ones not concerned with future states lead to outputs that are neither predictions nor projections (Fig. 1). Only the purpose to predict leads to a prediction as model output, whereas several other purposes that also look at potential future states (e.g. to explore, to inform management decisions) lead to projections. We can distinguish between these two types of outputs by looking at the boundary conditions: do we predict them with confidence and use these (either stationary or non-stationary) fixed in our model, or do we use them as a "control variable" [14] that we can experiment with to see what the effect may be on the system state (the "measure") (Fig. 1)?

As an example, weather models produce predictions because they assume boundary conditions about the state of the atmosphere for the coming days that they can (and do) predict with confidence [5]. In contrast, an agent-based model of pedestrian route choice produces projections of pedestrian densities in a city's streets for different route choice behaviors [15]. The boundary condition "percentages of pedestrians with route choice behavior A/B/C/D/E/F" are control variables that help to explore the state of city when different groups of people are roaming around. This is partly because we do not know how many pedestrians of a certain behavior type there are, or will be tomorrow, next week, or next year, but more importantly because a prediction for one of these moments is less

[1] We use the term theory in a loose sense; they can also be assumptions about a process or behavior that are not (yet) a theory, for example with the purpose to build such theory

interesting than the more general knowledge and understanding we derive from projections.

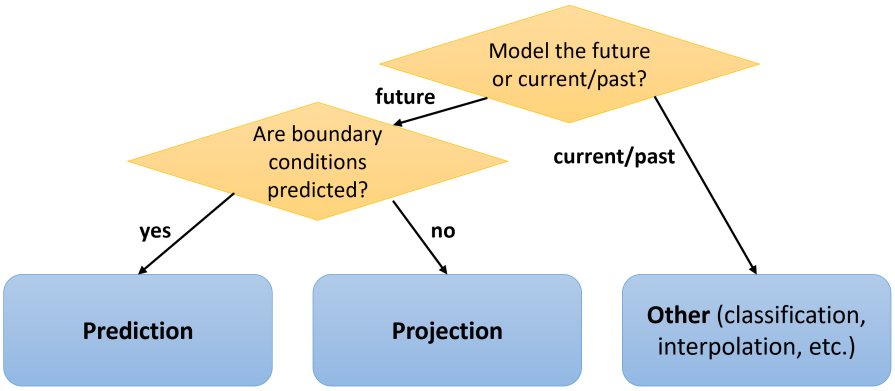

Fig. 1. Decision tree to catalogue model output as prediction, projection or other.

4 Spatial Machine Learning

"Prediction is very difficult, especially if it's about the future!" [16]

Whereas geosimulation models are formalizations of theories, spatial machine learning[2] models are algorithms trained on a dataset, either supervised or unsupervised. In machine learning, using the term prediction for a model output seems even more prevalent than in geosimulation, e.g., [2, 17, 18]. This is rather remarkable, given that most machine learning models are either used to fill in gaps at locations between or beyond observations, or to attach labels to unlabeled objects, based on one or more of their attributes. In both cases, the time interval of model output is thus the same as that of the observations or another period in the past for which we have object attributes. That is, neither of those are targeting claims about the future, and thus can be classified neither as predictions nor as projections given the definitions in section 2 (Fig. 1); we would rather call them *interpolations/extrapolations* or *classifications*, respectively.

It may well be the case that the (spatial) machine learning community applies a different definition of prediction that makes the term pertinent to data about current and past events too, such as "[to predict is to] anticipate well-defined aspects of data that are not currently known" [8]. Unknown aspects of data are

[2] We consciously use the term spatial machine learning and not GeoAI, as agent-based modelling (a geosimulation paradigm) is a technique stemming from AI, so using GeoAI would cause the two sets of methods in the previous and current chapter to confusingly overlap.

various and do not have to concern the future. We have not found this definition made explicit and we doubt that it reflects the intuitive perception of the term prediction by the general public. The discussion about model purposes that we saw in the geosimulation community seems not to have been fully taken up by the relatively young spatial machine learning community, although there are first signs of awareness. One example is the appearance of the term "explainable AI" (purpose to explain, instead of, or on top of, the purpose to classify or to predict, depending on the definition followed) [17]. Another is the recognition that a model does not have the same validity for spatial interpolation as for spatial extrapolation [18], which can be seen as a spatial analogue of modelling the past and the future in time. Such developments should be taken as an opportunity for the machine learning community to reconsider its model output terminology.

5 Conclusion

"All models are wrong, but some are useful" [19]

Even though the quote above appears in many of our lectures, we seem to forget about it when we term our model outputs "predictions". Many model outputs are *useful precisely because they are not predictions*, as the model purpose was, for example, to have a platform to enable discussions with peers or to inform decisions by assessing the effects of potential interventions (changing the boundary conditions of the system), rather than to make claims about what will happen in the future. The term prediction is overused.

Though this problem is not specific to geography, i.e. to geosimulation and spatial machine learning, resolving it requires us to be more precise in what purposes our modeling efforts have and how the purpose propagates to the model output. These purposes are specific for spatial and temporal information. Thus a more reflective use of terminology (and any change in general) has to start in our own backyards.

The climate science community has shown us that consistency in terminology is possible [5, 6]. It is important to be precise and nuanced about what our model outputs mean, because the models (and modellers) are judged by how well models accomplish what the beholder believes they should be doing. This is particularly crucial for model outputs, as opposed to other model components, because for the general public the model output is often the only model component they get to see; it is the component shown in news articles, reports, and public debates. Unclear terms for model outputs may thus unjustly reduce confidence.

Acknowledgements We would like to acknowledge two anonymous reviewers as well as everyone who ever asked us a question about model predictions and especially our colleague Michiel van Meeteren, who was the most recent crusader in this realm and made us realize that we now really need to write this down.

References

1. Berger, T., Troost, C.: Agent-based modelling of climate adaptation and mitigation options in agriculture. Journal of Agricultural Economics, **65**(2), 323-348 (2013), https://doi.org/10.1111/1477-9552.12045
2. Wang, J., Chen, R., He, Z.: Traffic speed prediction for urban transportation network: A path based deep learning approach. Transportation Research Part C: Emerging Technologies **100**, 372-385 (2019), https://doi.org/10.1016/j.trc.2019.02.002
3. Eigenbrod, F., Beckmann, M., Dunnett, S., Graham, L., Holland, R.A., Meyfroidt, P., Seppelt, R., Song, X.-P., Spake, R., Vaclavik, T., Verburg, P.H.: Identifying Agricultural Frontiers for Modeling Global Cropland Expansion. One Earth **3**(4), 504-514 (2020), doi10.1016/j.oneear.2020.09.006
4. Verstegen, J.A., van der Laan, C., Dekker, S.C., Faaij, A.P.C, Santos, M.J.: Recent and projected impacts of land use and land cover changes on carbon stocks and biodiversity in East Kalimantan, Indonesia. Ecological Indicators **103**, 563-575 (2019), https://doi.org/10.1016/j.ecolind.2019.04.053
5. MacCracken, M.: Prediction versus Projection–Forecast versus possibility. WeatherZine **26** (2001), https://sciencepolicy.colorado.edu/zine/archives/1-29/26/guest.html, last accessed 29 May 2023
6. IPCC Data Distribution Center: Guidance on the use of Data, Glossary. Available online: https://www.ipcc-data.org/guidelines/pages/definitions.html, last accessed 27 May 2023
7. Voros, J.: Big History and Anticipation. In: Poli, R.: Handbook of Anticipation, 1-40 (2017), Springer International Publishing, https://doi.org/10.1007/978-3-319-31737-3_95-1
8. Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montanola-Sales, C., Ormerod, P., Root, H., Squazzoni, F.: Different Modelling Purposes. Journal of Artificial Societies and Social Simulation **22**(3)6 (2019),https://doi.org/10.18564/jasss.3993
9. Epstein, J.M.: Why Model? Journal of Artificial Societies and Social Simulation **11**(4)12 (2008).
10. Cuddington, K., Fortin, M. J., Gerber, L. R., Hastings, A., Liebhold, A., O'connor, M., Ray, C.:. Process-based models are required to manage ecological systems in a changing world. Ecosphere, textbf4(2), 1-12 (2013)
11. Rounsevell, M. D., Robinson, D. T., Murray-Rust, D.: From actors to agents in socio-ecological systems models. Philosophical Transactions of the Royal Society B: Biological Sciences, **367**(1586), 259-269 (2012)
12. Rykiel, E.J.: Testing ecological models: the meaning of validation. Ecological Modelling **90**(3), 229-244 (1996). https://doi.org/https://doi.org/10.1016/0304-3800(95)00152-2
13. Troost, C., Huber, R., Bell, A.R., van Delden, H., Filatova, T., Le, Q.B., Lippe, M., Niamir, L., Polhill, J.G., Sun, Z., Berger, T.: How to keep it adequate: A protocol for ensuring validity in agent-based simulation. Environmental Modelling & Software **159**, 105559 (2023). https://doi.org/10.1016/j.envsoft.2022.105559
14. Sinton, D.: The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. In: Dutton, G., Harvard papers on geographic information systems, **7** (1978).
15. Filomena, G., Kirsch, L., Schwering, A., Verstegen, J. A.: Empirical characterisation of agents' spatial behaviour in pedestrian move-

ment simulation. Journal of Environmental Psychology **82**, 101807 (2022), https://doi.org/10.1016/j.jenvp.2022.101807
16. Bohr, N.: Prediction is very difficult, especially if it's about the future!" (1962), but date uncertain and quote often attributed to others, such as Mark Twain, Robert Storm Petersen, and Yogi Berra
17. Papadakis, E., Adams, B., Gao, S., Martins, B., Baryannis, G., Ristea, A.: Explainable artificial intelligence in the spatial domain (X-GeoAI). Transactions in GIS, **26**(6), 2413-2414 (2022), https://doi.org/10.1111/tgis.12996
18. de Bruin, S., Brus, D.J., Heuvelink, G.B.M, van Ebbenhorst Tengbergen, T., Wadoux, A.M.J-C.: Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics **69**, 101665 (2022), https://doi.org/10.1016/j.ecoinf.2022.101665
19. Box, G.E.P.: Science and statistics. Journal of the American Statistical Association, **71**(356), 791–799 (1976), https://doi.org/10.1080/01621459.1976.10480949. Note: "but some are useful" is not the original quote but a later expansion.